emerald PUBLISHING

ISSN 1468-4527
Volume **43** Number **1** **2019**

Online Information Review

ISSN 1468-4527

Guidelines for authors can be found at:
www.emeraldgrouppublishing.com/oir.htm

Online Information Review
is indexed and abstracted in:
ABI/Inform Complete
ABI/Inform Global
ABI/Inform Professional Advanced
ABI/Inform Professional Standard
Academic Research Library
Academic Search Alumni Edition
Academic Search Complete
Academic Search Premier
Cabell's
Chartered Association of Business Schools (CABS, UK)
CINAHL
CNKI
Compendex
CompuMath Citation Index
Computer Science Index
Computers & Applied Sciences Complete
Current Abstracts
Current Contents/Social and Behavioural Sciences
dblp Computer Science Bibliography
Discovery
Education Full Texts
Education Research Complete
EI Compendex
ERIC
Information Science & Technology Abstracts
INSPEC (Electrical & Electronics Abstracts)
ISI Alerting Service
Library and Information Science Abstracts
Library, Information Science and Technology Abstracts
Library Literature and Information Science Full Text
Library Literature and Information Science
Internet and Personal Computing Abstracts
Journal Citation Reports/Science Edition
Journal Citation Reports/Social Sciences Edition
OmniFile Full Text Mega
OmniFile Full Text Select
NAVIGA
Platinum Periodicals
Primo
Professional ABI/Inform Complete Professional ProQuest Central
ProQuest Advanced Technologies and Advanced Aerospace Collection
ProQuest Central
ProQuest Curriculum Essentials
ProQuest Computer Science Collection
ProQuest Education Journals
ProQuest Library Science
ProQuest Nursing and Allied Health Source
ProQuest Pharma Collection
ProQuest Sci Tech Journals
ProQuest Technology Journals
Readcube Discover
Research Library
Science Citation Index Expanded
Science Citation Index
Scopus
Social SciSearch
Summon
SWOCTS
Zetoc

Emerald Publishing Limited
Howard House, Wagon Lane,
Bingley BD16 1WA, United Kingdom
Tel +44 (0) 1274 777700; Fax +44 (0) 1274 785201
E-mail **emerald@emeraldinsight.com**

For more information about Emerald's regional offices please go to http://www.emeraldgrouppublishing.com/offices

Customer helpdesk:
Tel +44 (0) 1274 785278; Fax +44 (0) 1274 785201
E-mail **support@emeraldinsight.com**

Orders, subscription and missing claims enquiries:
E-mail **subscriptions@emeraldinsight.com**
Tel +44 (0) 1274 777700; Fax +44 (0) 1274 785201

Missing issue claims will be fulfilled if claimed within six months of date of despatch. Maximum of one claim per issue.

Hard copy print backsets, back volumes and back issues of volumes prior to the current and previous year can be ordered from Periodical Service Company.
Tel +1 518 537 4700; E-mail **psc@periodicals.com**
For further information go to **www.periodicals.com/emerald.html**

Reprints and permissions service
For reprint and permission options please see the abstract page of the specific article in question on the Emerald web site (**www.emeraldinsight.com**), and then click on the "Reprints and permissions" link. Or contact:
E-mail **permissions@emeraldinsight.com**

Emerald is a trading name of Emerald Publishing Limited

Printed by CPI Group (UK) Ltd, Croydon, CR0 4YY

EDITORIAL ADVISORY BOARD

Online Information Review
Vol. 43 No. 1, 2019
p. 1

1468-4527

Guest editorial

Social media mining for journalism

The exponential growth of social media as a central communication practice, and its agility in capturing and announcing breaking news events more rapidly than traditional media, has changed the journalistic landscape: social media has been adopted as a significant source by professional journalists, and conversely, citizens are able to use social media as a form of direct reportage. This brings along new opportunities for newsrooms and journalists by providing new means for newsgathering through access to a wealth of citizen reportage and updates about current affairs, as well as an additional showcase for news dissemination.

In addition to being a big opportunity and having changed the day-to-day practices in the newsrooms, social media has introduced a number of challenges when it comes to newsgathering, verification, production, reporting and dissemination. These include real-time monitoring of streams, event detection, noise filtering, contextualisation, source and content verification, fact checking, annotation and archiving. The development of more advanced algorithms and tools for journalists requires not only furthering research in computational techniques, but also engaging more closely with journalists to understand how they work, what problems they are facing when using social media and how their day-to-day workflows can be improved.

Social media are increasingly becoming the go-to platforms to get the news. A 2018 survey by the Pew Research Center found that as many as 62 per cent Americans use social media to get the news[1]. Likewise, news organisations are now employing full-time social media editors, and major news organisations such as Reuters[2] or the BBC[3] recommend their journalists to make frequent use of social media.

Research looking into social media use in journalism has also increased substantially in recent years. After Kwak *et al.*'s (2010) work highlighting the presence of news in social media, now cited over 5,000 times, an increasing number of works have studied social media as a platform that can be leveraged, *inter alia*, for researching, gathering and verifying breaking news (Diakopoulos *et al.*, 2012; Heravi and Harrower, 2016; Tolmie *et al.*, 2017; Zubiaga *et al.*, 2018; Konstantinovskiy *et al.*, 2018), for broadening the audience by maximising the diffusion of news (Diakopoulos and Zubiaga, 2014; McCollough *et al.*, 2017) or for news analytics (Castillo *et al.*, 2014; Zubiaga *et al.*, 2016).

This special section provides a gateway to look into a variety of research questions from both theoretical and practical perspectives. We highlighted four major topics of interest to the special section when we launched the call for papers. These include: newsgathering from social media, aiming to study algorithmic approaches to facilitating collection and research around newsworthy content with social media as a source (Zubiaga, 2018; Khare *et al.*, 2015; Heravi *et al.*, 2014); social media news analytics, where the objective is to analyse news readership from the perspective of social media as well as to perform additional analyses that give insights into how news circulates and is consumed online (Diakopoulos *et al.*, 2010); data and computational journalism, which aims to leverage social media data to enable computationally assisted production of journalistic content (Gray *et al.*, 2012; Heravi and McGinnis, 2015; Heravi, 2018); and ethics and digital citizenship, where ethical aspects of gathering eyewitness content

Online Information Review
Vol. 43 No. 1, 2019
pp. 2-6

1468-4527
DOI 10.1108/OIR-02-2019-395

This paper forms part of a special section "Social media mining for journalism".

from social media as well as other factors affecting diffusion of social media, such as gatekeeping and censorship, are explored (Frost, 2015).

This special section puts together seven articles covering these subjects. Given recent trends in the research area, the topics that prevail among these articles include verification of newsworthy content and detection of fake news, as well as event detection. In what follows, we provide brief summaries of these seven articles.

Opening up the discussion, the paper "A bibliometric analysis of event detection in social media" (Chen *et al.*, 2019) explores the research status and development trend of the field of event detection in social media through a bibliometric analysis of academic publications on Event Detection in Social Media research field between 2009 and 2017. The study suggests that the area of event detection in social media has received increasing attention and interest in academia with Computer Science and Engineering as two major research subjects. In terms of geographical contribution to the field, the paper identifies the USA and China to contribute the most to the publications on these topics. It further suggests that affiliations and authors researching this area tend to collaborate more with those within the same country. Finally, the paper identifies 14 research themes in this area, as part of which a number of newly emerged themes, such as Pharmacovigilance event detection, are discovered.

The paper titled "What the fake? Assessing the extent of networked political spamming and bots in the propagation of #fakenews on Twitter" (Al-Rawi *et al.*, 2019) uses 14m tweets with #fakenews posted from January 3 to May 7, 2018, investigating what and who are behind in promoting and propagating fake news as a national issue. In particular, the authors examine: the most associated users and hashtags mentioned in #fakenews tweets, and the most active Twitter accounts in spamming and disseminating the #fakenews tweets. A large portion of the tweets were attacks against CNN and other mainstream media outlets, which is partly a result of the success of networked political spamming by conservative groups. Investigating the most active users promoting #fakenews tweets, it turned out the majority of the most active accounts likely came from spamming bots. This study has provided insight into Twitter users' networked spamming accounts that influenced the discussion on fake news on Twitter.

In another paper, titled "A corpus of debunked and verified user-generated videos" (Papadopoulou *et al.*, 2019), the authors built an annotated data set, called Fake Video Corpus 2018 (FVC-2018), of 380 user-generated videos that contain 200 debunked (fake) and 180 verified (real) videos uploaded in YouTube, Facebook and Twitter. The data set also contains 77,258 tweets that shared any of the 380 videos. The authors followed the definition of the fake videos proposed in (Teyssou *et al.*, 2017) and extended the initial Fake Video Corpus data set compiled in the same study. In addition to the efforts to build the annotated data set, the authors also provide a detailed analysis of the descriptive statistics of the videos and helped to understand the characteristics of the data set. The FVC-2018 data set provides a valuable resource for a challenging benchmark and future studies on video verification.

The study in the paper titled "Location impact on source and linguistic features for information credibility of social media" (Aladhadh *et al.*, 2019) investigates the impact of location on information source and credibility level in social media with tweets of a diverse set of events across multiple countries. In particular, the authors examine: the types of sources expected in different events from both in- and outside the country of events, and linguistic features among sources of different type, credibility level, and location. The authors found that the distribution of some sources differs between locations significantly and the tweets of the same credibility level have different linguistic features based on their distance from an event and the topic of an event. The results of this study provide insights for improving current credibility models when applied to

different domains: most importantly, such models need to be trained on data from the same place of event.

In their paper on "Event news detection and citizens community structure for disaster management in social networks", Toujani and Akaichi (2019) present a methodology that combines the detection of natural hazards from social media with determination of endangered communities as a result of those natural hazards. They present a methodology which consists of three steps: first, they perform a set of natural language processing methods to detect event triggers, extract named entities mentioned in those texts, which helps identify the communities and areas involved in the natural hazards, and a dependency analysis which is intended to mitigate the ambiguity of social media posts; second, the apply fuzzy techniques on the extracted events to cluster related posts; and, third, these clusters are leveraged in order to determine communities which are at risk owing to the effects of the natural hazard. The authors show the effectiveness of their methodology with experiments on 26 crisis events as well as a set of synthetic data sets, outperforming other baselines. The paper also shows a tool that enables visualisation of the events and communities identified by the system. The tool is intended to facilitate, among others, journalists' work of sifting through large collections of tweets posted during these natural hazards.

In the following paper, the authors explored the potential of NodeXL as a tool for analysis and visualisation in the context of news diffusion. This is the paper titled "Social media analytics: analysis and visualisation of news diffusion using NodeXL" (Ahmed and Lugovic, 2019), where the authors first conducted a comprehensive literature review and showed how effective NodeXL is to understand reactions in social media. NodeXL, for example, can discover the most shared URLs, popular hashtags, or influential users from the stream of social media posts, and such features are helpful for newsrooms to cover social media. As NodeXL is easy to use without any programming language, journalists also can easily include social media content to their stories and potentially attract more online readers by showing how online communities react to certain topics.

Wrapping up the articles in this special section, Chio takes an exploratory approach to the quantification of journalistic values. In the paper titled "An exploratory approach to the computational quantification of journalistic values", the author matches the textual indices extracted through automated content analysis, with human conceptions of journalistic values which were derived from surveying journalism grad students (Choi, 2019). The results of this paper suggest that the numbers of words and quotes news articles contain have a strong association with the survey respondent assessments of their balance, diversity, importance and factuality. Additionally, the paper suggests that the assessment of journalistic values influences the perception of news credibility. In terms of specific indicators, the paper suggests that linguistic polarisation is an inverse indicator of respondents' perception of balance, diversity and importance. While linguistic intensity was shown to be useful for gauging respondents' perception of sensationalism, the paper suggests that it is an ineffective indicator of importance and factuality. Furthermore, the number of adverbs and adjectives in news articles appear to be useful for estimating respondents' perceptions of factuality and sensationalism. Finally, the study suggests that the greater the numbers of quotes, pair quotes and exclamation/question marks in a news headline, the lower the respondents' perception of journalistic values in that news article would be.

The papers presented in this special section illustrate the extensiveness and potentials of social media mining for journalism and news industry, including news and event detection, analytics, verification and journalistic values associated, and/or affected by, the use of social media in this domain.

We, the guest editors, would like to extend our appreciation to the authors who submitted to this special section, as well as the reviewers who dedicate their time to furthering the

research of our contributors. We are looking forward to the continued growth and evolution of this rapidly growing interdisciplinary field of research.

Arkaitz Zubiaga
Department of Computer Science, The University of Warwick, Coventry, UK
Bahareh Heravi
School of Information & Comms Studies, University College Dublin, Dublin, Ireland
Jisun An
Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar, and
Haewoon Kwak
Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

Notes

1. www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/
2. http://handbook.reuters.com/index.php?title=Reporting_From_the_Internet_And_Using_Social_Media
3. http://news.bbc.co.uk/1/shared/bsp/hi/pdfs/26_03_15_bbc_news_group_social_media_guidance.pdf

References

Ahmed, W. and Lugovic, S. (2019), "Social media analytics: analysis and visualisation of news diffusion using NodeXL", *Online Information Review*, Vol. 43 No. 1, pp. 149-160.

Aladhadh, S., Zhang, X. and Sanderson, M. (2019), "Location impact on source and linguistic features for information credibility of social media", *Online Information Review*, Vol. 43 No. 1, pp. 89-112.

Al-Rawi, A., Groshek, J. and Zhang, L. (2019), "What the fake? assessing the extent of networked political spamming and bots in the propagation of# fakenews on twitter", *Online Information Review*, Vol. 43 No. 1, pp. 53-71.

Castillo, C., El-Haddad, M., Pfeffer, J. and Stempeck, M. (2014), "Characterizing the life cycle of online news stories using social media reactions", *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 211-223.

Chen, X., Wang, S., Tang, Y. and Hao, T. (2019), "A bibliometric analysis of event detection in social media", *Online Information Review*, Vol. 43 No. 1, pp. 29-52.

Choi, S. (2019), "An exploratory approach to the computational quantification of journalistic values", *Online Information Review*, Vol. 43 No. 1, pp. 133-148.

Diakopoulos, N. and Zubiaga, A. (2014), "Newsworthiness and network gatekeeping on twitter: the role of social deviance", *ICWSM*, pp. 587-590.

Diakopoulos, N., De Choudhury, M. and Naaman, M. (2012), "Finding and assessing social media information sources in the context of journalism", *Proceedings of the Sigchi Conference on Human Factors in Computing Systems*, pp. 2451-2460.

Diakopoulos, N., Naaman, M. and Kivran-Swaine, F. (2010), "Diamonds in the rough: social media visual analytics for journalistic inquiry", *2010 IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pp. 115-122.

Frost, C. (2015), *Journalism Ethics and Regulation*, Routledge.

Gray, J., Chambers, L. and Bounegru, L. (2012), *The Data Journalism Handbook: How Journalists can Use Data to Improve the News*, O'Reilly Media.

Heravi, B. (2018), "Data journalism in 2017: a summary of results from the global data journalism survey", in Chowdhury, G., McLeod, J., Gillet, V. and Willett, P. (Eds), *Transforming Digital Worlds*, Springer International Publishing, pp. 107-113.

Heravi, B. and Harrower, N. (2016), "Twitter journalism in Ireland: sourcing and trust in the age of social media", *Information, Communication & Society*, Vol. 19 No. 9, pp. 1194-1213.

Heravi, B. and McGinnis, J. (2015), "Introducing social semantic journalism", *The Journal of Media Innovations*, Vol. 2 No. 1, pp. 131-140.

Heravi, B., Morrison, D., Khare, P. and Marchand-Maillet, S. (2014), "Where is the news breaking? Towards a location-based event detection framework for journalists", *Multimedia Modeling*, Springer International Publishing, pp. 192-204.

Khare, P., Torres, P. and Heravi, B. (2015), "What just happened? A framework for social event detection and contextualisation", *2015 48th Hawaii International Conference on System Sciences*, pp. 1565-1574, 10.1109/HICSS.2015.190.

Konstantinovskiy, L., Price, O., Babakar, M. and Zubiaga, A. (2018), "Towards automated factchecking: developing an annotation schema and benchmark for consistent automated claim detection", preprint arXiv:1809.08193.

Kwak, H., Lee, C., Park, H. and Moon, S. (2010), "What is twitter, a social network or a news media?", *Proceedings of the 19th International Conference on World Wide Web*, pp. 591-600.

McCollough, K., Crowell, J.K. and Napoli, P.M. (2017), "Portrait of the online local news audience", *Digital Journalism*, Vol. 5 No. 1, pp. 100-118.

Papadopoulou, O., Zampoglou, M., Papadopoulos, S. and Kompatsiaris, I. (2019), "A corpus of debunked and verified user-generated videos", *Online Information Review*, Vol. 43 No. 1, pp. 72-88.

Teyssou, D., Leung, J.-M., Apostolidis, E., Apostolidis, K., Papadopoulos, S., Zampoglou, M., Papadopoulou, O. and Mezaris, V. (2017), "The invid plug-in: web video verification on the browser", *Proceedings of the First International Workshop on Multimedia Verification*, pp. 23-30.

Tolmie, P., Procter, R., Randall, D.W., Rouncefield, M., Burger, C., Wong Sak Hoi, G., Zubiaga, M. and Liakata, M. (2017), "Supporting the use of user generated content in journalistic practice", *Proceedings of the 2017 Chi Conference on Human Factors in Computing Systems*, pp. 3632-3644.

Toujani, R. and Akaichi, J. (2019), "Event news detection and citizens community structure for disaster management in social networks", *Online Information Review*, Vol. 43 No. 1, pp. 113-132.

Zubiaga, A. (2018), "Mining social media for newsgathering", preprint arXiv:1804.03540.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M. and Procter, R. (2018), "Detection and resolution of rumours in social media: a survey", *ACM Computing Surveys (CSUR)*, Vol. 51 No. 2, p. 32.

Zubiaga, A., Liakata, M., Procter, R., Hoi, G.W.S. and Tolmie, P. (2016), "Analysing how people orient to and spread rumours in social media by looking at conversational threads", *PloS One*, Vol. 11 No. 3, p. e0150989.

What municipal websites supply and citizens demand: a search engine optimisation approach

Carlos Serrano-Cinca
Faculty of Economy and Business Administration, University of Zaragoza, Zaragoza, Spain, and
Jose Felix Muñoz-Soro
Aragonese Foundation for Research and Development, Aragon, Spain

Received 4 February 2018
Revised 23 April 2018
12 July 2018
Accepted 17 July 2018

Abstract

Purpose – The purpose of this paper is to analyse if citizens' searches on the internet coincide with the services that municipal websites offer. In addition, the authors examine municipal webpage rankings in search engines and the factors explaining them.
Design/methodology/approach – The empirical study, conducted through a sample of Spanish city councils, contrasted if the information that can be found on a municipal website fits with citizens' demands. This has been done by comparing the most-searched keywords with the contents of municipal websites.
Findings – A positive relationship between the supply and demand of municipal information on the internet has been found, but much can still be improved. Analysed administrations rank the basic data of the organisation, as well as some of the fundamental competences thereof, at the top in search engines, but the results are not entirely effective with some keywords still highly demanded by citizens, such as those related to employment or tourism. Factors explaining internet ranking include the number of pages of the municipal website, its presence in social networks and an indicator designed to measure the difficulty of ranking the municipal place-name.
Originality/value – The results obtained from this study provide valuable information for municipal managers. Municipal websites should not only include information in which citizens are interested, but achieve accessibility standards, have a responsive web design, and follow the rules of web usability. Additionally, they should be findable, which also requires improvement in terms of the design of the municipal website thinking in search engines, particularly in terms of certain technical characteristics that improve findability. A municipal website that wants to have a good positioning should increase its contents and attain the maximum degree possible of visibility in social networks.

Keywords Search engine optimisation, e-government, Internet ranking, Municipal websites
Paper type Research paper

1. Introduction

Public administrations offer information and online services to citizens through their websites, improving their transparency (Tirado-Valencia *et al.*, 2016), allowing interaction on social networks (Gandía *et al.*, 2016), boosting the semantic web (Muñoz-Soro *et al.*, 2016) and even enhancing e-democracy (Fietkiewicz *et al.*, 2017). The number of citizens that connect to municipal websites is becoming greater and greater; in fact, municipal websites are the most common channel through which citizens communicate with governmental agencies (Ebbers *et al.*, 2016). In the European Union, the percentage of individuals using the internet for interaction with public authorities is about 48 per cent, with a maximum of 88 per cent in Denmark (Eurostat, 2017a, b). But doubts have arisen when attempts have been made to determine if citizens' demands coincide with what municipal websites offer, and also if municipal information is easy to find through the use of search engines. The main objective of this paper is to contribute to solving these doubts; and, to this end, three research questions are tackled.

The analysis reported in this paper was supported by Grant ECO2013-45568-R from the Spanish Ministry of Education and Science and the European Regional Development Fund and by Grant Ref. S-14 (3) and S-86 from the Government of Aragon.

This paper forms part of a special section "Social media mining for journalism".

Online Information Review
Vol. 43 No. 1, 2019
pp. 7-28

1468-4527
DOI 10.1108/OIR-02-2018-0042

Testing how local governments respond to their constituents' priorities should be an important focus of research (Einstein and Kogan, 2016). Hence, the first research question analyses if that for which citizens search coincides with what municipal webpages offer. Perhaps citizens are worried about issues like job searches, leisure possibilities, or the procedure necessary to procure a grant; however, often such information does not appear on the website, or, if it does appear, it is not easy to find, even through the use of internet search engines. Municipal websites, especially those of a small size, usually exhibit deficiencies in many aspects (Gandía and Archidona, 2008; Feeney and Brown, 2017). This first area of research is explored by means of "keyword tools", resources provided by main search engines regarding the search average conducted for a keyword in a determined period of time.

Search engine optimisation (SEO) aims to achieve that a given website appears in the first position when a search involving specific keywords is employed by users (Evans, 2007). The study of Baye *et al.* (2016), which analysed 12,000 search terms used by 2m users on top retailers, confirms the importance of carrying out SEO actions, because most users find a website due to a search engine. Although there are no equivalent studies applying to municipal websites, it seems reasonable to think that, if city councils offering services to citizens want to attract tourists to their municipality or simply want to publish their history, they cannot only focus their activities on the creation of content or the design of questionnaires, but also have to worry about the positioning of their websites in search engines. Therefore, the second research question aims to analyse the positioning level of municipal websites. In order to answer this question, it is possible to simply introduce the keywords that best define the municipal competences into a search engine and then write down the positions they occupy. Although it is a laborious activity, it is possible to automate this process using SEO tools.

The third research question has to do with the factors that explain internet positioning. Although the algorithms employed by the main search engines are not known in detail, the fundaments of the one used by Google were published in a conference (Page and Brin, 1998), and some empirical studies have analysed factors that determining the positioning of websites (Baye *et al.*, 2016; Moreno and Martinez, 2013; Evans, 2007; Luh *et al.*, 2016; Su *et al.*, 2014). According to such studies, a website's contents are a key aspect for search engines, both in terms of quantity and quality. Thus, it is expected that municipalities with a solid website, one in which both competences are reported and online procedures are possible, will rank top in search engines. Other hypotheses have to do with the extent of an organisation's social networks presence, the receiving of backlinks – links from other websites – and the difficulty of positioning certain terms that have very high competition. The identification of factors that favour positioning could help policymakers to improve municipal websites – such factors including transparency, citizen participation, the interoperability of the information, accessibility, usability and findability.

This paper makes various contributions. Numerous studies have analysed municipal websites, especially the disclosure of financial information (Armstrong, 2011; da Cruz *et al.*, 2016), the use of social networks (Bonsón *et al.*, 2012), or their accessibility (King and Youngblood, 2016). This research proposes a methodology to contrast if the information available on a municipal website fits with citizens' demands, comparing the most-searched keywords with the contents of municipal websites. We consider that findability should be one of the quality attributes of a municipal website, a line of investigation in which the work of Kopackova *et al.* (2010) on Czech municipalities stands out. The use of SEO tools is proposed in order to monitor the internet positioning of municipal websites. Furthermore, two indicators have been designed in order to measure the difficulty of positioning a municipal website, starting with the number of webpages indexed by Google as belonging to the public administration, the number of inhabitants of the municipality, and the number of results obtained when searching the place-name of the municipality in Google. Finally, the drivers of positioning have been analysed – an aspect that has previously been studied for online shops (Su *et al.*, 2014), but which we apply to the case of municipal websites.

The empirical research was conducted via a sample of Spanish city councils. A positive link between the supply and demand of municipal public information on the internet has been found, but much can still be improved. Therefore, it is recommended that municipal managers consult the words most-searched by citizens and then redirect the contents of their municipal website. It has been observed that public administrations rank the basic data of their organisation and some basic competences very well in search engines; however, the results are often not so good for keywords that are highly demanded by citizens, such as those related to employment or tourism. Municipal managers should adopt strategies to improve their website's positioning in search engines, which will result in a better service for citizens. They must consider that, as decisive factors of internet positioning, the number of pages that the municipal web has, the municipality's presence in social networks, and the indicators that measure the difficulty of positioning the municipal website stand out.

The rest of the research is organised as follows. Section 2 presents the literature review and the hypothesis development. Section 3 displays the empirical study. Finally, policy recommendations and conclusions are presented.

2. Literature review and hypothesis

Numerous empirical studies have analysed the content of municipal websites (Serrano-Cinca *et al.*, 2009; Armstrong, 2011; Bonsón *et al.*, 2012; Gandía *et al.*, 2016; Brusca *et al.*, 2016, among others). Alcaide-Muñoz *et al.* (2017) perform a meta-analytic review of such studies, concluding that the transparency of information about governments depends on both institutional and environmental factors. The size of the municipality stands out among the factors that explain the level of disclosure of public information (Ferraz Esteves de Araujo and Tejedo-Romero, 2016), but transparency is also best achieved when the government has no absolute majority (García-Sánchez *et al.*, 2013). Among the explanatory factors of municipal transparency, Cuadrado-Ballesteros *et al.* (2014) have studied political ideology and rivalry, and Caba-Pérez *et al.* (2014) have studied political competence. The percentage of the dependent population is also an explanatory factor in the case of information on sustainability (Alcaraz-Quiles *et al.*, 2015). The interaction of city councils with citizens through the use of social networks has also been studied; however, this a subject that still has a long way to go (Gandía *et al.*, 2016). Finally, e-democracy, which includes aspects such as electronic voting (Fietkiewicz *et al.*, 2017), is the last stage of e-government, according to a meta-analytic review of e-government by Lee (2010), which describes ten e-government stage models, ranging from stage 1 (simple information on the website), to stage 10 (e-democracy).

In addition to the content of municipal websites, the technological and formal aspects thereof have also been studied. Peristeras *et al.* (2009) review the use of the semantic web in public administrations. They find that the public administration domain still lacks commonly agreed-upon content standards, in spite of the advantages that the reutilisation of public data provides, especially due to the tendency of introducing open-government data programs (García-Tabuyo *et al.*, 2017). Accessibility stands out among the formal aspects of municipal websites that try to ensure that any person, including those with limitations, can browse municipal websites without a problem (King and Youngblood, 2016). Additionally, citizens are mostly connected to the internet via mobile devices, so municipal websites must have a responsive web design. Finally, the usability of municipal websites is important in terms of improving citizens' experiences and the impression that cities make on third-interest parties (Youngblood and Mackiewicz, 2012).

However, findability, another important aspect, has been studied much less than other factors. Findability is about making information easy to find; Hedden (2008, p. 1) claims: "if it cannot be found, it may as well not exist". Findability requirement could be considered as part of accessibility requirements (White, 2003). Websites must be easy for citizens to use,

but must also be designed to consider the characteristics that search engines value, such as employing friendly URLs, avoiding the use of rich media files and iframes, and adding relevant keywords in the meta description tag of the website (Su *et al.*, 2014). According to Kopackova *et al.* (2010), no regulation may command of search engines that municipalities' webpages must be shown in the first search position; findability depends upon a municipality's goodwill.

The first research question analyses if that which citizens demand coincides with the information municipal websites offer. According to Bearfield and Bowman (2017), community demand plays an important role in fostering municipal transparency. In the same way, it also seems reasonable to think that, in municipalities with a high percentage of internet users, the demand for municipal information will be high. Thus the local government would be expected to answer such demand through the provision of a wide variety of content and services on the municipal website (Serrano-Cinca *et al.*, 2009). High users of technologies of information and communication (ICTs) are more positive in terms of their attitudes towards e-government (Gauld *et al.*, 2010). However, McNeal *et al.* (2003) found that e-government implementation is driven by legislative professionalism and, to a lesser extent, state professional networks, rather than citizen demands. In the previously mentioned studies, the percentage of households with internet access is used as a proxy for public demand for e-government services, while, in our study, we analyse keywords used by the citizens in search engines, providing a more direct and complete approach. If the municipal website's objectives are aligned with citizens' searches, there will be a positive correlation between the searched words and the words that appear in the municipal website; therefore, the following hypothesis is proposed:

H1. A positive relationship between citizens' searches and the words that appear in the municipal website is expected.

In contrast to accessibility requirements, which are maintained in accordance with the law, no findability requirements are stated for e-government websites, and no findability tests are held (Kopackova *et al.*, 2010). In the same way that local governments are usually located in emblematic buildings and in a privileged part of the city, it should also be deemed as desirable that they take up a distinguished place in the results of search engines too, especially in terms of searches related to their competences. Some of the factors that explain internet positioning are well known (Baye *et al.,* 2016; Evans, 2007; Luh *et al.*, 2016; Su *et al.*, 2014). A wide array of content is a key aspect that search engines value; in fact, experts in online marketing claim that "content is the king". Google clearly affirms this position. Since the update of its algorithms in October 2014, websites with wide content have improved their positioning (Chandra *et al.*, 2015). Therefore, city councils with wide and well-documented municipal websites, which not only develop contents about municipal services but also other issues (such as tourist references of the locality, hiking routes, historical occurrences, or distinguished figures born in the locality), should reach a better internet positioning. Consequently, the following hypothesis is proposed:

H2. A positive relationship between the size of the municipal website and its ranking in search engines is expected.

Beyond the amount of information, the quality of a municipal website is measured using different attributes, such as the relevance, value-added, timeliness, completeness or accessibility. Academic literature has collected numerous indexes designed to evaluate the quality of a municipal website (Armstrong, 2011; da Cruz *et al.*, 2016; Miranda *et al.*, 2009). Search engines give more importance to websites that provide information than to commercial websites, which only exist to sell products (Chandra *et al.*, 2015). City councils

that own high-quality websites and accomplish accessibility standards should reach a better internet ranking. Consequently, the following hypothesis is proposed:

H3. A positive relation between the quality of the municipal website and its ranking in search engines is expected.

In contrast to the first generation of municipal websites, where citizens were limited to the passive viewing of content, a 2.0 website allows both parties to interact by means of social networks. Bonsón *et al.* (2012) studied 2.0 websites in city councils, finding that their use is related to previous e-government levels of development. Gandía *et al.* (2016) found an essential ornamental focus in 2.0 web implementation: city councils are more focused on promotional issues than on the disclosure of information about the entity's management. Search engines value the social aspects of webpages, such as the possibility of sharing contents; in fact, social media optimisation has become one of the emerging SEO activities. It seems reasonable to think that city councils should also have a good positioning in social networks; therefore, the following hypothesis is proposed:

H4. A positive relationship between the municipality's social networks presence and its ranking in search engines is expected.

The study of web impact factors has a long pedigree (Ingwersen, 1998; Ortega *et al.*, 2014; Vaughan, 2014). Generally, if a website has many backlinks, it is an important website. Hence, web search engines use backlink counting as a means to measure the importance of a website (Page and Brin, 1998). Usually, websites with sufficient content of high-quality receive links from other websites, but sometimes SEO practitioners also carry out strategies to increase backlinks in order to improve a website's positioning. Consequently, the following hypothesis is proposed:

H5. A positive relationship between backlinks to the municipal website and its ranking in search engines is expected.

Due to the competition that can exist, and taking into account the fact that there are many websites focused on the same subject, not all terms show the same difficulty when positioning, so the difficulty is even higher if competitors have important websites on the internet or, in Google's terminology, websites with a high page rank (Page and Brin, 1998). The empirical results of the study by Luh *et al.* (2016) reveal that page rank is still a dominant factor in Google's ranking algorithm. It should also be considered that domain authority is highly correlated with search engine rankings (Mavridis and Symeonidis, 2015), which gives city councils' websites an advantage. Domain authority is calculated by evaluating multiple factors into a single score, including linking root domains and the number of total links; hence, sites with a very large number of high-quality external links are at the top end of the domain authority scale (Zhang and Cabage, 2017). For online shops, keyword research software is employed to identify online niches; namely, searches with a high-demand, but low-supply (Wilson and Pettijohn, 2008). In the case of a city council, the positioning of words related to employment, sport, or tourism is more complicated than the positioning of words related to the municipal competence of cemeteries. Moreover, the municipality's name may cause additional difficulty. For example, Castelserás and Andorra are two neighbouring Spanish communities, but there is also a country called Andorra. According to Google, there are 285,000,000 websites containing the word "Andorra", compared to the 181,000 with the word "Castelserás", so the managers of Andorra's municipal website have a huge handicap when seeking to position their terms. Consequently, the following hypothesis is proposed:

H6. A negative relationship between the difficulty of positioning the place-name and its positioning in search engines is expected.

3. Empirical study

We posited the possibility of conducting a cross-national study with major European cities, as with Bonsón *et al.* (2012), but this proposal was discarded for four reasons: first language is very relevant to this study, as it affects the searched terms in Google; second due to social and cultural reasons, what citizens seek may vary from country to country; for example, it seems reasonable to assume that, for a country with a high unemployment rate, the number of searches for municipal job offers would be high; third the study focuses on the competences assigned to local administrations, which vary from country to country; and forth the results of an analysis carried out with a sample of large town halls cannot be extrapolated to small town halls, which also serve a considerable percentage of the population. In fact, the rural population in Europe constitutes 22.8 per cent of the total population (Eurostat, 2017a, b).

Therefore, a cross-national study was discarded and we chose to analyse a single country (Spain). We also thought about conducting the study via an analysis of the largest Spanish town halls, as in Serrano-Cinca *et al.* (2009). However, four languages are spoken in Spain and controlling that effect would have been very complicated – we would have had to search in Catalan, Galician and Basque, as well as in Spanish, which could have affected search engine positioning. Moreover, as has been said, we thought that it would be especially interesting to analyse the search engine positioning of small town councils, which usually face the greatest difficulties in technological terms (Feeney and Brown, 2017). This is why we chose to focus on a single Spanish region, Aragón, selecting a sample that included data from large, medium and small municipalities. Geographically located in the northeast of Spain, Aragon is a region that is usually considered trustworthy in terms of representing the Spanish average. Aragón has 731 municipalities, of which 708 have less than 5,000 inhabitants. All 23 city councils that have more than 5,000 inhabitants were chosen, and also another 33 smaller than those were chosen through random sampling. The study sample consisted of 56 municipal websites, which represent 90.14 per cent of Aragon's population. The study was performed in June 2017 (all the obtained data are available upon request).

Table I shows the variables used in the study. In the first place, a web content analysis was performed on the municipal websites. To this end, a questionnaire with 98 questions – inspired by Serrano-Cinca *et al.* (2009), Armstrong (2011), Bonsón *et al.* (2012) and da Cruz *et al.* (2016) – was designed. Table II shows the 98 questions and the range of values.

Variable	Definition
Q-index	Quality index of the municipal website, obtained adding 98 variables pertaining to organisation, transparency, participation, compulsory competences, non-compulsory competences and 14 formal aspects about the accessibility and usability of the website
SITE	Number of webpages that the municipal website contains, measured using the command "site:" in the Google search engine
SOC-index	The active and updated municipal presence in six social networks (Twitter, Facebook, LinkedIn, YouTube, Instagram and WhatsApp) and the use of five tools that improve the social use of the web (wikis, blogs, podcast, mashups and RSS)
BACKLINK	Link popularity. Number of incoming links according to WooRank, an SEO service that is available for free
KE-index	Keyword effectiveness index: square number of monthly searches of the place-name divided by the number of results obtained when you search the place-name in a search engine
PLACE-EASY	Place-name Easy ratio: number of inhabitants of the municipality divided by the number of results obtained when the place-name is introduced in a search engine
SHARE	Internet Market Share ratio: number of pages of the municipal website divided by the total results that are obtained when the place-name is introduced in the search engine
POSIT	Municipal website positioning, obtained recording the position that the website obtains for 98 keywords, adding them and taking the reciprocal

Table I.
Variables employed for the hypothesis testing and their definition

Question	Search term	Number of searches	1st position (%)		1st page (%)	
Organisation						
The municipality has a webpage (0/1)	*ayuntamiento*	40,500	48	85.7	52	92.9
Contact information (0/1)	*sede electrónica*	60,500	12	21.4	36	64.3
Municipal regulations (0/1)	*reglamento*	2,400	18	32.1	30	53.6
Public job offers (0/3)	*oposiciones*	1,000,000	9	16.1	26	46.4
Contact form (0/1)	*teléfono*	49,500	28	50.0	39	69.6
Catalogue of procedures (0/1)	*oficina virtual*	22,200	7	12.5	17	30.4
Citizens can check the status of the procedures (0/1)	*trámites y servicios*	2,400	20	35.7	35	62.5
General request form (0/3)	*solicitud*	2,900	26	46.4	36	64.3
Electronic notifications (0/1)	*notificaciones*	110,000	14	25.0	26	46.4
Bulletin board (0/1)	*tablón de anuncios*	40,500	13	23.2	28	50.0
Verification point (0/1)	*código de verificación*	1,900	2	3.6	7	12.5
Citizen's charter and/or management indicators (0/1)	*carta de servicios*	320	9	16.1	20	35.7
Buyer profile (0/1)	*perfil del contratante*	4,400	25	44.6	45	80.4
Processing of public tenders (0/3)	*licitaciones*	2,900	15	26.8	34	60.7
Information about electronic invoice (0/1)	*factura electrónica*	6,600	10	17.9	24	42.9
Processing and payment of taxes (0/3)	*impuestos*	2,900	19	33.9	32	57.1
Processing of tax claims (0/3)	*recursos*	2,400	10	17.9	28	50.0
Transparency						
Transparency portal (0/1)	*transparencia*	3,600	17	30.4	27	48.2
Political organisational chart (0/1)	*organigrama*	9,900	17	30.4	26	46.4
Decisions on interest conflicts (0/1)	*resolución de conflictos*	1,300	9	16.1	15	26.8
Salaries of councillors and employees (0/1)	*sueldo alcalde*	70	5	8.9	14	25.0
Date and agenda of the next plenaries (0/1)	*pleno*	880	20	35.7	21	37.5
Minutes of councils (0/1)	*actas*	1,000	20	35.7	25	44.6
Videos of councils available online (0/1)	*video pleno*	10	4	7.1	16	28.6
Budget for the current year (0/3)	*presupuesto*	6,600	15	26.8	24	42.9
Budget settlement (0/3)	*liquidación*	2,900	8	14.3	23	41.1
Accounts for the current year (0/3)	*contabilidad*	8,100	2	3.6	18	32.1
Audit reports (0/1)	*auditoría*	4,400	8	14.3	16	28.6
Debt reports (0/1)	*deuda*	1,000	6	10.7	23	41.1
Average period of payment to suppliers (0/1)	*periodo medio de pago*	390	21	37.5	26	46.4
Inventory of goods and/or vehicles (0/1)	*inventario*	2,400	6	10.7	23	41.1
Offers to the public tenders (0/1)	*contratos*	2,400	18	32.1	30	53.6
Small amount contracts (0/1)	*contratos menores*	590	8	14.3	18	32.1
Agreements, with the amount and beneficiaries (0/1)	*convenio*	3,600	18	32.1	30	53.6
Subsidy calls, with the criteria (0/1)	*subvenciones*	5,400	27	48.2	39	69.6
Judgments that affect the entity (0/1)	*sentencias*	1,000	5	8.9	19	33.9
Processing of public information requests (0/3)	*información*	301,000	13	23.2	32	57.1
Participation						
Active discussion forums (0/1)	*foro*	22,200	2	3.6	16	28.6
Online surveys (0/1)	*encuestas*	12,100	9	16.1	24	42.9
Active chats (0/1)	*chat*	246,000	4	7.1	7	12.5
e-voting (0/1)	*voto por Internet*	2,400	5	8.9	11	19.6
Online participation in the plenaries or councils (0/1)	*participación*	1,300	20	35.7	27	48.2
Suggestions allowed in the budget preparation (0/1)	*presupuestos participativos*	480	8	14.3	16	28.6

(*continued*)

Table II. Questionnaire with the keywords, monthly search average, and the city councils that appear in first place and page

Question	Search term	Number of searches	1st position (%)		1st page (%)	
Compulsory competences						
Processing of changes of address (0/3)	*cambio de domicilio*	720	12	21.4	21	37.5
Processing of census certificates (0/3)	*certificado de empadronamiento*	6,600	14	25.0	21	37.5
Processing of partnership certificates (0/3)	*certificado convivencia*	1,300	13	23.2	19	33.9
Information about urban planning (0/3)	*PGOU*	2,900	22	39.3	27	48.2
Processing of building permits (0/3)	*licencia de obras*	590	18	32.1	23	41.1
Processing of urban information certificates (0/3)	*urbanismo*	2,400	24	42.9	30	53.6
Information about local history and heritage (0/1)	*historia*	27,100	15	26.8	33	58.9
Procedures about housing of public promotion (0/3)	*VPO*	1,600	4	7.1	12	21.4
Procedures about housing rehabilitation (0/3)	*rehabilitación vivienda*	90	8	14.3	15	26.8
Information about parks and gardens (0/1)	*parque*	12,100	7	12.5	27	48.2
Information about pollution and noise (0/1)	*contaminación*	5,400	4	7.1	21	37.5
Information about incidents in public services (0/1)	*avería*	1,300	5	8.9	15	26.8
Information about water quality (0/1)	*calidad del agua*	170	13	23.2	18	32.1
Procedures about water supply (0/3)	*agua*	33,100	15	26.8	26	46.4
Procedures about wastewater management (0/3)	*alcantarillado*	480	6	10.7	21	37.5
Street map (0/1)	*mapa*	135,000	4	7.1	24	42.9
Reservation of municipal facilities (0/3)	*equipamientos*	140	16	28.6	31	55.4
Information of social need situation (0/3)	*servicios sociales*	4,400	19	33.9	36	64.3
Procedures for immediate attention for people in situations of social exclusion risk (0/3)	*asistente social*	1,900	24	42.9	34	60.7
Claim and payment of fines (0/3)	*policía*	40,500	23	41.1	33	58.9
Complaints e-mailbox (0/1)	*multas*	5,400	9	16.1	25	44.6
Procedures about civil protection (0/3)	*protección civil*	6,600	7	12.5	22	39.3
Procedures about fire prevention and extinction (0/3)	*bomberos*	14,800	15	26.8	28	50.0
Up-to-date traffic info (0/1)	*trafico*	49,500	1	1.8	17	30.4
Public transport information (0/1)	*autobuses*	33,100	7	12.5	22	39.3
Information on urban mobility for disabled people (0/1)	*minusválido*	1,000	7	12.5	22	39.3
Geographical information about the city (0/1)	*como llegar*	201,000	10	17.9	23	41.1
Touristic offer (0/1)	*turismo*	9,900	9	16.1	29	51.8
Vacation activities (0/1)	*vacaciones*	40,900	1	1.8	8	14.3
Procedures to participate in fairs, markets, etc. (0/3)	*mercado*	9,900	12	21.4	28	50.0
Procedures about garbage collection and street cleaning (0/3)	*basura*	4,400	24	42.9	29	51.8
Procedures about dog census (0/3)	*perros*	165,000	3	5.4	11	19.6
Procedures about cemeteries and funeral services (0/3)	*cementerio*	5,400	12	21.4	34	60.7
Calendar with local festivities (0/1)	*calendario*	165,000	5	8.9	28	50.0
Agenda with municipal activities (0/1)	*agenda*	40,500	27	48.2	36	64.3
Consulting the catalogue of municipal libraries (0/1)	*biblioteca*	27,100	8	14.3	34	60.7
Reservations for the use of sports facilities (0/3)	*deporte*	40,500	25	44.6	38	67.9

Table II. (*continued*)

Question	Search term	Number of searches	1st position (%)		1st page (%)	
Register for municipal activities and courses (0/3)	*cursos*	14,800	12	21.4	27	48.2
Information about school centres (0/1)	*colegio*	33,100	6	10.7	18	32.1
Information about ICTs (0/1)	*TIC*	18,100	10	17.9	20	35.7
Non-compulsory competences						
Directory of services (0/1)	*servicios*	6,600	26	46.4	35	62.5
News about city council and the city (0/1)	*noticias*	1,500,000	12	21.4	32	57.1
Meteorological information (0/1)	*el tiempo*	6,120,000	1	1.8	6	10.7
Information on Agenda 21 and similar environmental initiatives (0/1)	*agenda 21*	2,900	17	30.4	31	55.4
Information about business and employment (0/1)	*empleo*	4,090,000	10	17.9	24	42.9
Information about industrial land (0/1)	*empresas*	9,900	5	8.9	14	25.0
Information about development cooperation (0/1)	*ONG*	14,800	8	14.3	22	39.3
Procedures about infant and adult education (0/3)	*educación*	18,100	22	39.3	33	58.9
Procedures about associations (0/3)	*asociación*	6,600	11	19.6	27	48.2
Health-related services (0/3)	*cita previa*	550,000	2	3.6	10	17.9
Consumer protection information (0/3)	*oficina del consumidor*	8,100	14	25.0	24	42.9
Information on gender violence (0/1)	*violencia de genero*	18,100	13	23.2	24	42.9
Services for elderly care (0/3)	*cuidado de ancianos*	12,100	4	7.1	9	16.1
Procedures for promote equal opportunities (0/3)	*becas*	49,500	14	25.0	29	51.8
Request of birth, marriage and death certificates (0/3)	*registro civil*	27,100	2	3.6	21	37.5

Notes: $n = 56$. The first column shows the questions analysed, and the range of values. The 0-3 scale is used to assess procedures and services. The values that can be adopted are: 0, if the procedure is not present on the web; 1, when only website-based information is provided about it; 2, when it is possible to download forms; and 3, when it is permitted to submit completed forms and payments, if necessary. The second column shows the 98 keywords that have been analysed by category, but in Spanish. The monthly search average, obtained using Google's Keyword Planner, is shown in the third column. The next columns display the number of city councils that appear in first place in the search results and on the first page, in both absolute values and in percentage

Table II.

The variables were classified into five categories: organisation, transparency, participation, compulsory competences and non-compulsory competences. The first category, municipal organisation, includes aspects such as ways in which the administration can be contacted, if service letters are published, the availability of the buyer profile or if the payment of taxes can be conducted online. Transparency includes indicators like the publication of the organisation chart, municipal salaries, information about grants and agreements or municipal financial information. Participation includes aspects such as the existence of electronic voting or participatory budgets. Spanish law establishes 15 compulsory municipal competences, which are: a register of inhabitants, urbanism, urban environment, waters, infrastructures, social assistance, security, mobility, tourism, commerce, public health, cemeteries, free time, education and encouragement in the use of ICTs. Finally, non-compulsory municipal competences include benefits provided in order to offer a better service to the municipality's citizenry, for example, in services related to health, consumption claims, or meteorological information. In addition to the 98 questions about contents, 14 formal aspects about accessibility and usability were included, leading to 112 variables, which were added in order to obtain a global quality index of the website (Q-index).

The next variable (SITE) tries to measure the quantity of information a municipal website contains. For that purpose, we use the command "site:" of the Google search engine, which counts the number of webpages of the municipality indexed by the search engine. For example, if we write into Google "site:zaragoza.es", it informs us that the search engine has indexed 1,060,000 webpages from Zaragoza's City Council. If we write "tourism site:zaragoza.es", it lets us know that 39,800 pages of the website containing the word tourism have been indexed.

Third, in order to measure a municipality's presence in social networks, an index inspired by Gandía *et al.* (2016) was designed, which takes into account 11 aspects (SOC-index). The SOC-index captures the active and updated municipal presence in six social networks (Twitter, Facebook, LinkedIn, YouTube, Instagram and WhatsApp) and the use of five tools that improve the social use of the web (wikis, blogs, podcast, mashups and RSS). The number of followers for each network and the number of posts were recorded, as well as the date of the last post. A value of zero was assigned to the municipalities that did not have any followers or that did not provide updated information. Fourth, the link popularity was measured via our recording of the number of received links (BACKLINK). Given the difficulty in obtaining the number of links to a website, the use of web mentions as accurate substitutes for inlinks has been proposed – an aspect analysed by Ortega *et al.* (2014). We compared three SEO tools (WooRank, Majestic and Alexa), and we were also able to contrast the results with data provided by the managers of some of the municipal websites. Both Majestic and WooRank showed a very high correlation. Ad Alexa's data differed significantly, it was discarded. WooRank was chosen as it is a recognised SEO tool that has been used in various academic works (Mistry *et al.*, 2013; Ferraz, 2015).

The next indicators measure the facility to position a city council in a search engine. One widely employed method is based on the keyword effectiveness index (KE-index). There are several ways to calculate the KE-index (Wilson and Pettijohn, 2008); we chose to use Wylie's (2012) definition (p. 258). The KE-index is computed by obtaining "the number of monthly searches of the keyword" and squaring it; then dividing this result by "the number of results obtained when introducing the keyword in the search engine". In our case, the keyword is the municipal place-name:

$$\text{Keyword Effectiveness Index} = \frac{\text{Number of monthly searches of the keyword}^2}{\text{Total of results obtained when introducing the keyword in the search engine}}.$$

The numerator "Number of monthly searches of the keyword" comes from Google's Keyword Planner (Figure 1), but other services, such as KWFinder.com or Semrush.com, can also be used. The Google Keyword Planner can be configured in such a way that it only targets searches conducted in a specific location. This is a very useful option for municipal managers; nevertheless, it should be noted that although citizens are the most regular users of such webpages, tourists also use them. We considered that it was a good choice not to use the local filter for the empirical study, since analysed a set of municipalities, not a single municipality. The denominator "Total of results obtained when introducing the keyword in the search engine" comes from Google. In all the searches carried out in Google, the personalised search options were deactivated.

We designed two new indicators to measure the difficulty faced by municipal websites in terms of search engine rankings. The first one is calculated by dividing the number of inhabitants of the municipality by the total of the results obtained when the place-name is introduced in the search engine. We called it the Place-name Easy ratio (PLACE-EASY). The smaller a municipality is, and the greater the results obtained when the place-name is introduced in the search engine, the higher the difficulty of positioning will be. The second indicator is obtained by dividing the number of pages of the municipal website by the total of the results obtained when introducing the place-name in the search engine. This indicator

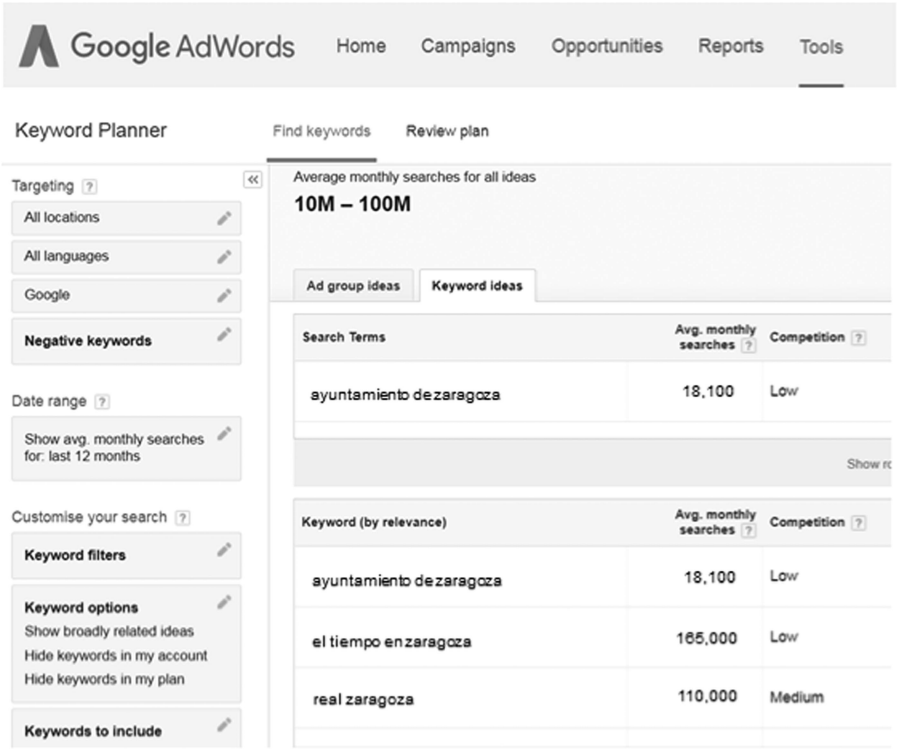


Figure 1.
Screenshot of the Keyword Planner tool, from Google, which allows the number of monthly searches for specific terms to be known

shows which part of the total presence of the place-name on the internet corresponds to the city council's website. For this reason, we named it the internet market share ratio (SHARE).

Finally, the dependent variable measures the positioning of the municipal website (POSIT). Google is able to obtain personalised search results by analysing cookies, the use of a Google account and the IP address. The computer used to perform the search could introduce biases by analysing cookies, the traces of Google search history and Google account, and IP location. But you can prevent Google from taking personalised search results into account by turning Google personal results off, deleting cookies and preventing the browser from sending location data following the guidance provided by Google Help documentation. In any case, the software used to analyse the search engine positioning, Rank Checker, offers an option to remove personalisation bias (Figure 2). Nevertheless, we recognise that it is possible that Google's local search capability could have biased the results in some way. Another possibility would be to use

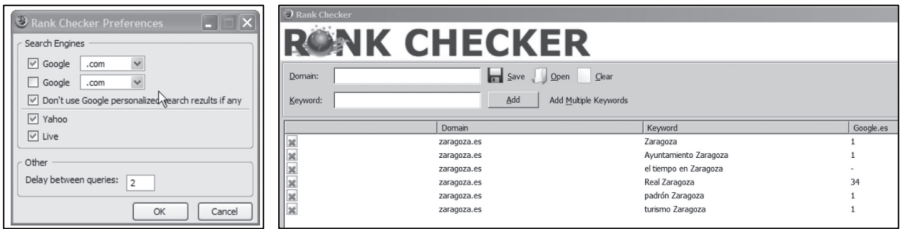


Figure 2.
Screenshots of the Rank Chequer tool, of SEObook, which allows researchers to calculate a search engine's ranking of a website from a list of keywords

an anonymous proxy server to hide the IP address when browsing the Web. An alternative approach would be to search each municipality using a device located in the same municipality or tools that simulate this. This would allow us to take into account the way in which each municipality uses local SEO, which is increasingly important. Nonetheless, considering that factors such as the IP address could influence the results, the fieldwork was carried out in a way that minimised biases. The methodology achieved this in two ways. First, all the municipalities chosen for analysis are located in a limited geographical space – they all belong to the same region, Aragón – and the differences between search results grow as physical distances between the locations of users increase (Kliman-Silver *et al.*, 2015). Second, the searches were conducted by the two authors in parallel in the two main cities of Aragón – Zaragoza and Huesca – and the results obtained were identical for all the municipal websites.

Municipal websites may compete locally and in a large geographical area. As an example of the latter, a tourist municipality can compete with other municipalities to attract tourists, trying to rank their offer in search engines with respect to other municipalities and using terms such as "the best Mediterranean beach". However, for other terms, the competition is local, competing with private companies, other administrations, or even NGOs. An example is the competition between municipal and private schools, or a municipal cemetery vs a private cemetery. The approach followed in the paper is that municipal websites mostly compete locally and, therefore, we used the search term in combination with the city's place-name (e.g. school + Zaragoza). Even in the case of tourism, we could assume that the person who searches is already interested in the specific city. As a result, competition among municipalities was not analysed, but the competition that takes place between the website of the municipality and other websites that deal with the same term.

For each of the 98 items about content in the global quality index of the website (Q-index), the most-searched word by citizens was identified through the use of the Google Keyword Planner. Table II shows the list with the 98 chosen keywords and the number of monthly searches. Fieldwork was carried out in order to verify the fulfilment of the 98 items of the questionnaire for each of the municipal websites. We decided that the two authors should complete the questionnaire separately, and the results obtained were very similar for both authors. The most common cause of mismatch was that one of the authors was not able to find the required information on the municipal website. In these cases, it was simply confirmed together. The questionnaire was subsequently delivered to those responsible for the municipal websites of the major city councils, who subsequently contacted the authors to inform them that some of the items did exist. In this way, we made sure that the data was consistent. After this, all keywords were analysed, one by one, adding the place-name of the municipality and recording the positions they had. When the municipal web did not appear in the first 200 results of the search engine, the number 200 was assigned. Later, the positions obtained by each word were added. The greater the value of the obtained variable, the worse the positioning of the municipal website; thus, the variable was transformed by taking the reciprocal.

3.1 What citizens search for on the internet

Table II shows that the most-searched terms related to municipal competences have to do with tourism, sport, mobility, and traffic states. Terms related to other competences, such as cemeteries, exhibited significantly fewer enquiries. Terms like how to go to attained 201,000 monthly searches, while cemetery reached 5,400. Among non-compulsory municipal competences, public jobs reached 1,000,000; news 1,500,000; and weather 6,120,000.

In order to know about the municipal information offered, the number of pages of the municipal websites that include the 98 keywords have been counted, one at a time, using the "site:" command of Google's search engine. If the municipal website's objectives are aligned with citizens' searches, there will be a positive correlation between the searched words and the words that appear in the municipal website. For each municipal website, the Spearman

correlation coefficient between both variables was calculated. In 40 of 56 analysed city councils (71 per cent), the correlation coefficient was positive and statistically significant; the average being 0.339, which is both positive and statistically significant. To sum up, a positive relationship between citizens' searches and the words that appear in the municipal website was found within the analysed data (*H1*).

Figure 3 displays a scatter plot, with the X axis showing a ranking of the words searched by citizens and the Y axis showing a ranking of the words that municipal websites offer. Absolute values have been transformed to their ranks in order to help display the results. Four quadrants are shown: the upper-left quadrant shows words that appear in websites but are not often used by citizens, like plenary, participation, or equipment. The upper-right quadrant shows the words most searched for by citizens that appear frequently on the web, such as information, sport, news, or job. The lower-right quadrant displays words that are used frequently by citizens, but do not appear much on the website, such as public jobs, chat, appointment, and domestic violence. Municipal managers should make a stronger effort to take such subjects into account on their websites. Finally, the fourth quadrant, lower-left, shows words with a low number of searches that do not appear much on the website, like mayor salary or participatory budgets.

3.2 SEO analysis

In order to analyse the search engine positioning of municipal websites, the ranking that each city council has in Google for each of the 98 analysed keywords has been noted.

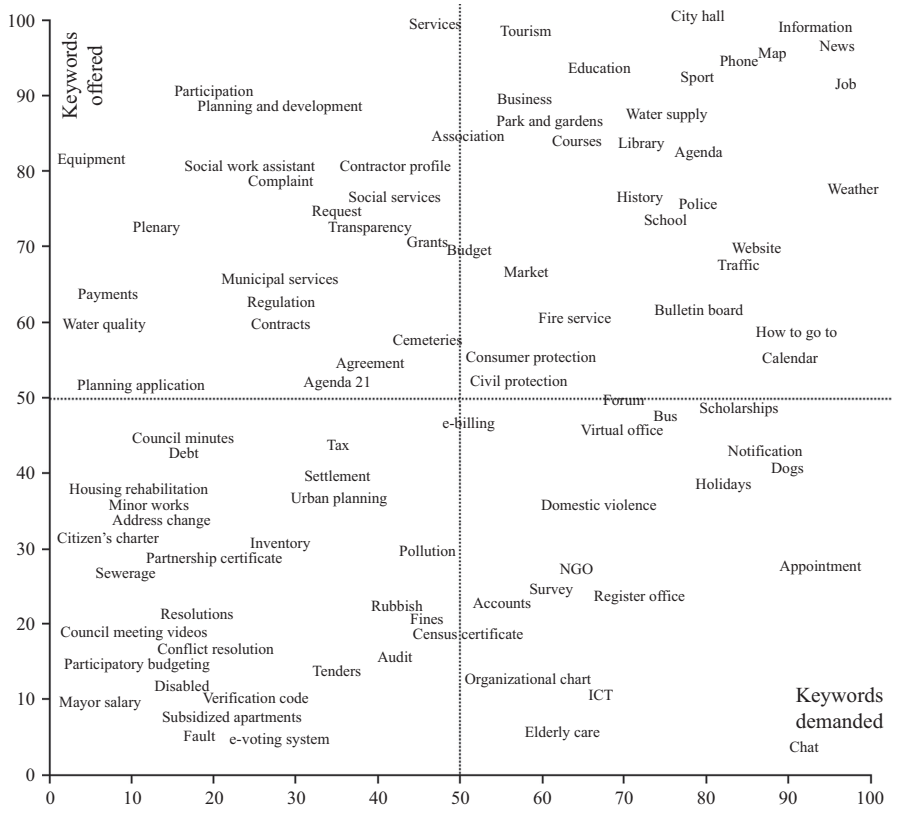


Figure 3.
Supply and demand of municipal information: terms used in municipal websites against terms searched by citizens

Table II shows the words that rank the city councils as better or worse. The number of city councils that appear in the first place of the search results for each word has been counted, as well as those on the first results page, which means among the first ten positions. In terms of average values, municipal websites appear in the first position of the search engine in 22.18 per cent of searches for the 98 analysed keywords. In 43.88 per cent of cases, these websites appear among the first ten positions. Among individual results, Zaragoza, the capital of Aragon, with 700,000 inhabitants, ranks in first position for 50 of the 98 keywords (51.02 per cent), and ranks on the first page of Google for 82 of the 98 keywords (83.67 per cent). This is an indicator that municipalities should monitor in their dashboards.

By analysing the table, it is observed that, in general, city councils rank the basic data of their own organisation very well. For example, if we search for X city council, the city council's website will probably appear in the first place; this is the case for 48 of the 56 analysed city councils (85.7 per cent). The fact that city councils manage to rank important terms related to their compulsory competences correctly, like grants, urban planning, buyer profile, or social work assistant, is remarkable. Greater difficulties are found for terms like job or tourism, for which 16 per cent of the administrations analysed hardly manage to appear in the first place. However, these are very competitive words, so by applying SEO techniques, and with a great effort, councils could appear in the first position of search engines.

Search engine ranking is, in general, quite adequate for sections such as municipal organisation and competences, while transparency and participation are the categories with the worst rankings. In these two categories, the lack of municipal experience in matters like participatory budgets or e-voting system is added to the difficulty of ranking competitive terms like forum or accounts. Many terms that have to do with non-compulsory competences inhere serious positioning difficulties – for example, terms like holidays, in which just one city council was positioned in first place (ahead of travel agency portals).

Figure 4 shows a scatter plot, with the X axis displaying the ranking of the most-searched words by citizens, and the Y axis showing the ranking of the better-positioned words of municipal websites. It is possible to identify highly searched words by citizens that are not well positioned, like elderly care, chat, weather, appointment, or domestic violence. Similar figures constructed with data from their websites could be useful for municipal managers to detect the words that citizens use that are not well positioned.

3.3 Factors that explain the search engine positioning of municipal websites

In this epigraph, the hypotheses about the determinant factors explaining the internet positioning of municipal websites are empirically contrasted. In Table III, which shows the Pearson linear correlation coefficient between independent variables, a positive statistically significant relation is observed between quantity and quality (with a Pearson correlation coefficient of 0.565), as well as between quantity of information and backlinks (with a coefficient of 0.655). Additionally, the correlation between social networks presence and the Quality Index stands out, being 0.606. Regarding the indexes of difficulty, there is not much correlation. The only appreciable result comes from the KE-index ratio and the quality index, which is 0.430.

A regression taking search engine positioning as the dependent variable was also performed, employing the previous ones as explanatory variables. Table IV presents the results of nine models. The first seven models are univariant models and show each independent variable separately. Number 8 is the full model, and number 9 is a parsimonious model. All univariate models present statistically significant values, except model number 5, which enters KE-index. A positive relation between the size of the municipal website and its ranking in search engines was found within the analysed data (*H2*). A positive relationship between the quality of the municipal website and its ranking in search engines was also found (*H3*), along with a positive relationship between the municipality's social networks presence and its ranking in search engines (*H4*), and a

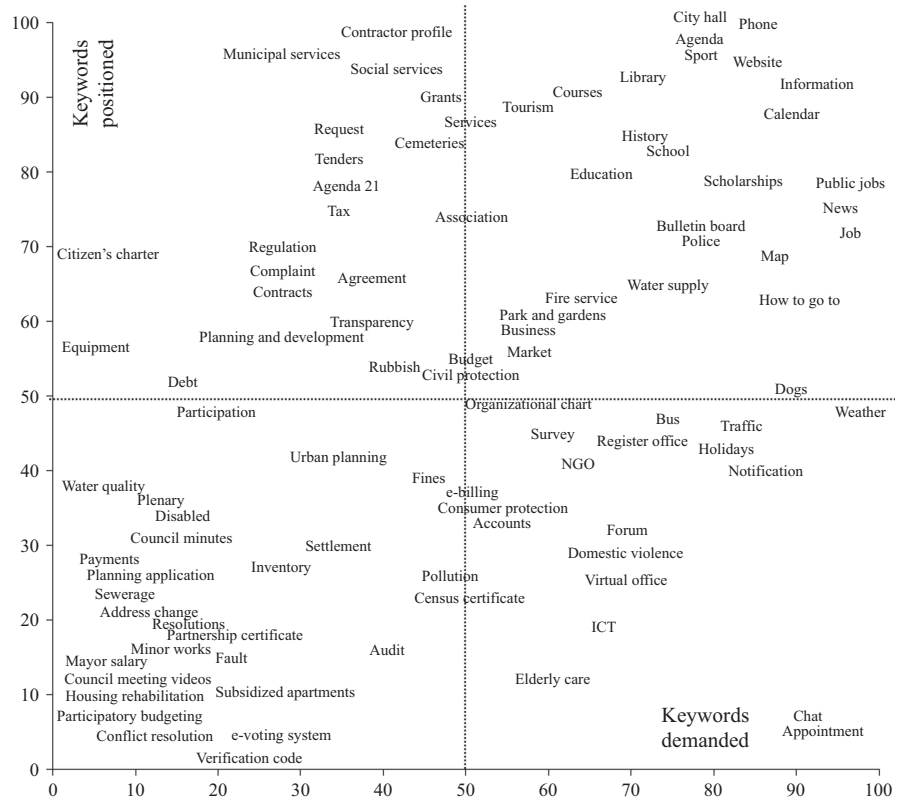


Figure 4.
SEO and demand of municipal information: municipal keywords positioned in search engines against terms searched by citizens

	SITE	Q-index	SOC-index	BACKLINK	KE-index	PLACE-EASY	SHARE
Number of pages (SITE)	1	0.565**	0.323*	0.655**	0.187	−0.038	0.110
Quality (Q-index)		1	0.606**	0.401**	0.430**	−0.119	0.133
Social network (SOC-index)			1	0.385**	0.358**	0.073	0.337*
Received links (BACKLINK)				1	0.240	−0.082	−0.050
Keyword effectiveness index (KE-index)					1	−0.090	0.110
Place-name Easy ratio (PLACE-EASY)						1	0.139
Municipal internet Market Share ratio (SHARE)							1

Notes: *,**Significant at 0.05 and 0.01, respectively

Table III.
Pearson's correlation coefficients among the determinant factors of positioning in search engines ($n = 56$)

positive relationship between backlinks to the municipal website and its ranking in search engines was found within the analysed data (*H5*). To test *H6* on the difficulty of positioning in search engines, three variables were employed: Place-name Easy ratio (PLACE-EASY), Municipal internet Market Share ratio (SHARE), and KE-index. A negative relationship between the difficulty of positioning the place-name and its positioning in search engines was found within the analysed data (*H6*) via the use of PLACE-EASY and SHARE, but not

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Number of pages (SITE)	0.789**							0.927**	0.715**
Quality (Q-index)		0.694**						0.154	
Social network (SOC-index)			0.575**					0.257*	0.252**
Received links (BACKLINK)				0.717**				−0.323	
Keyword Effectiveness Index (KE-index)					0.281			−0.054	
Place-name Easy ratio (PLACE-EASY)						0.366*		0.300**	0.304**
Municipal Internet Market Share ratio (SHARE)							0.401**	−0.031**	
Adjusted R^2	0.614	0.420	0.315	0.502	0.057	0.114	0.141	0.806	0.807
F	69.388**	32.166**	20.797**	44.392**	3.588	6.516*	8.046**	26.504**	60.828**

Notes: $n = 56$. The table shows estimated standardized coefficients of the regression. *,**Significant at 0.05 and 0.01, respectively

Table IV.
Regression that exhibits the explicative factors of positioning (POSIT)

with the KE-index. To sum up, according to the results of the study, all proposed hypotheses are accepted within the analysed data.

The full model (number 8) presents an adjusted R-square of 0.806, but some of the coefficients do not have the expected sign. Also, Table III reveals evidence of multicollinearity between some of the independent variables. These factors lead to some of the model's variables being discarded. In this way, we obtained the last model: a parsimonious model with a small number of independent variables, all of which present the greatest possible explanatory power. Using too many variables could lead to overfitting of data; hence, parsimonious models are useful to ensure validity. Another reason is to reduce the cost of measuring all of the variables for prediction purposes. Automatic model selection procedures could be used to obtain parsimonious models, but they have many limitations and have been criticised because they do not incorporate the judgement of experts (Huberty, 1989). Gelman and Hill (2007, p. 69) recommend starting with a simple model before building in additional complexity. Our choice was to generate a set of reasonable models, all of which rely on subject-matter expertise, and to compare their goodness of fit and their prediction accuracy, based on Akaike's Information Criterion (Akaike, 1974), Bayesian Information Criterion (Schwarz, 1978), and the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996). LASSO shrinks the regression coefficients towards zero, and many of them to exactly zero, hence resulting in parsimonious models. Nowadays, these are the most-often used methods to obtain parsimonious models and, in addition, they follow systematic procedures. The same results were obtained using the three procedures. Three variables designed to explain municipal web search engine positioning entered the parsimonious model: the number of pages on the municipal website, the municipal's presence in social networks, and the Place-name Easy ratio. The adjusted R^2 presents a value of 0.807, which indicates a good adjustment.

3.4 Future research lines

As an attribute related to the quality of the municipal website, we propose further investigation into the design of the management indicators for scorecards that help the

managers of public administrations to track website findability. De Andrés *et al.* (2010) found a positive relationship between accessibility and findability, since meeting accessibility guidelines leads to a mark-up of the structural elements, which, in turn, increases the findability of the website. But the relationship between findability and other attributes that characterise the online information of municipal websites is not so clear. Hence, new studies can be conducted regarding the attribute findability and its inclusion in models of structural equations designed to analyse its relationship with the other attributes of municipal webpages, such as credibility, trustworthiness, authority, accessibility, or usability, among others. Municipal websites may rank well for some keywords and not so well for the rest, and it would be interesting to analyse the differences between the rankings as well as the explanatory factors. We are considering conducting future work which will analyse ranking differences among city councils' websites and the possible factors that influence these. The analysis of the competition between different municipalities within a specific geographical area and with respect to a specific search term such as "the best Mediterranean beach" would be also of interest. Finally, it would be interesting to adapt studies, such as the one by Tavares and Da Cruz (2017), on factors that affect transparency to the case of findability, in order to determine whether findability is driven by leadership, capacity and other political traits of local governments, or, rather, whether it hinges on other external factors.

4. Policy recommendations

Some of the first academic approaches to website evaluation included searchability, along with content, authority, web organisation and accessibility amongst the criteria (Collins and Flick, 1995). Findability, as an attribute, is considered to be very important for online shops (Constantinides, 2002), but not so important with regard to the information provided by municipal websites – with some exceptions, like Kopackova *et al.* (2010). Municipal websites must not just achieve accessibility standards, but also have a responsive web design, and follow the rules of web usability. Additionally, we recommend they should be findable, which requires revisions to the design of municipal website thinking in search engines, particularly with regard to certain technical characteristics that improve findability.

Huang and Benyoucef (2014) established a positive link between usability and credibility. It is also possible to think that there exists a positive link between findability and credibility, since a website appearing in a prominent search engine place could be perceived by citizens as a sign of authority. According to Margetts (2009), governments are trying to influence social behaviour and shape the outside world through the use of four basic tools: nodality, authority, treasure and organisation. Nodality denotes the property of being "nodal" to information and social networks and having the capacity to both disseminate and collect information. She concludes that nodality is crucially affected by the decisions of the most popular search engines; thus, the nodality of a government will depend upon that government's ability to compete successfully in the online space. Hence, improving the findability of a municipal web not only provides a better service to citizens, but also promotes other important aspects, like the credibility of the municipality and its communication with citizens. In fact, city councils are often placed physically in emblematic buildings, usually in a prestigious city-centre location. However, in many cases, counsel websites have an internet presence that does not fit with their importance. If, whilst promoting their institutional credibility, local administrations want to provide services to their citizenry, attract tourists to the municipality, or publish their history and heritage, they should put effort into ensuring their websites rank correctly in search engines.

Finally, municipal website objectives must be aligned with the interests of citizens, which involves that municipal websites should clearly tend towards a user-oriented design. But how should municipalities obtain users feedback? Cegarra-Navarro *et al.* (2012) address the

need for local administrations to discover citizens' needs in real time, suggesting the use of groupware systems formed by both civil servants and citizens. Feeney and Brown (2017) describe some initiatives carried out by large municipalities, but they recognise that the building of this type of evolving, adaptable and responsive approach to e-government requires flexibility, innovation, and enormous resources, which many smaller cities may not have. In order to know citizens' needs, our proposal is that municipalities should analyse citizens' internet search data, something that all municipalities can do at a low cost.

5. Conclusions

This paper analyses if citizens' searches on the internet coincide with the services that municipal websites offer. In addition, we examine municipal webpage rankings in search engines and the factors explaining them. We propose identifying the keywords that best define municipal competences, analysing the keywords most searched by citizens, and studying website positioning in search engines in order to help local administrations to use SEO tools. The empirical study was conducted with a sample of Spanish city councils. These rank the basic data of their own organisations pretty well, along with important terms related to their competences, but they experience difficulties with terms like job or tourism. On average, in the analysed sample, municipal websites were able to rank 22.18 per cent of the 98 analysed terms in first place and 43.88 per cent amongst the first ten positions – that is, on the first page of the search engine. In the same way that online shops do not settle simply for showing their catalogue to customers and waiting for them to enter, municipal managers should make efforts to ensure their websites rank highly on the principal search engines, so that citizens can quickly find their services. It is proposed that local administrations set their own positioning objectives, and that they should add relevant SEO key performance indicators to their balanced scorecard.

According to the results of the study, two of the factors explaining search engine positioning are the number of pages that municipal websites have and their social networks presence. Therefore, a municipal website that wants to have a good positioning should increase its contents and attain the maximum possible visibility in social networks. Via this method, aside from providing citizens with a better service, the number of backlinks received will increase, a factor directly related to internet positioning. It is obvious that bad praxis, whose only objective consists of cheating search engines, must be avoided. These actions are called black hat SEO practices, and when they are detected, search engines punish their use.

Some city councils have a handicap if their place-name coincides with the name of a country, a region, another municipality, a surname or a common word. For example, there are dozens of cities in several countries that are called Washington, besides being both the name and surname of persons, as well as the name of higher education institutions and infrastructure projects throughout the world. These cities face additional difficulties in terms of appearing in the top positions of a search engine. We propose two indicators that measure the extrinsic difficulties that affect the positioning in search engines, the most relevant being the Place-name Easy ratio, which is obtained by dividing the number of inhabitants of the municipality by the total results obtained when introducing the place-name in the search engine. We have found a negative relationship between the difficulty of positioning the place-name and its ranking in search engines.

Moreover, the use of SEO tools, like the ones used in this research, can provide municipal website managers with guidance regarding the interests of citizens, as well as enabling a comparison between such interests and to the content of the municipal website in question. In this study, we have compared whether the information that can be found in municipal websites fits with citizens' demands. In 71 per cent of the analysed city councils, the correlation coefficient between citizens' searched keywords and what the municipal

websites offered was both positive and statically significant. It was observed that citizens often search for terms like elderly care, chat, weather, appointment, or domestic violence, but that such terms are not very often in municipal websites. Thus, the municipal managers should strive to include these issues on their websites. It should be noted that an SEO analysis should be done for each city council, because localisms are important. For example, in a touristic municipality, the search demand will focus on monumental routes, local festivities, weather, maps, postal codes, how to travel by bus or pictures of the locality. Finally, the fact that in an autonomous region of Spain many of the city councils' websites display the specific content their citizens demand cannot be extrapolated to other contexts. The use of one sample from one country, instead of multiple samples from different countries, is a limitation of the paper, which encourages new empirical studies, using samples from different countries.

References

Akaike, H. (1974), "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, Vol. 19 No. 6, pp. 716-723.

Alcaide-Muñoz, L., Rodríguez-Bolívar, M.P. and López-Hernández, A.M. (2017), "Transparency in governments: a meta-analytic review of incentives for digital versus hard-copy public financial disclosures", *The American Review of Public Administration*, Vol. 47 No. 5, pp. 550-573.

Alcaraz-Quiles, F.J., Navarro-Galera, A. and Ortiz-Rodríguez, D. (2015), "Factors determining online sustainability reporting by local governments", *International Review of Administrative Sciences*, Vol. 81 No. 1, pp. 79-109.

Armstrong, C.L. (2011), "Providing a clearer view: an examination of transparency on local government websites", *Government Information Quarterly*, Vol. 28 No. 1, pp. 11-16.

Baye, M.R., De los Santos, B. and Wildenbeest, M.R. (2016), "Search engine optimization: what drives organic traffic to retail sites?", *Journal of Economics & Management Strategy*, Vol. 25 No. 1, pp. 6-31.

Bearfield, D.A. and Bowman, A.O.M. (2017), "Can you find it on the web? An assessment of municipal e-government transparency", *The American Review of Public Administration*, Vol. 47 No. 2, pp. 172-188.

Bonsón, E., Torres, L., Royo, S. and Flores, F. (2012), "Local e-government 2.0: social media and corporate transparency in municipalities", *Government Information Quarterly*, Vol. 29 No. 2, pp. 123-132.

Brusca, I., Manes Rossi, F. and Aversano, N. (2016), "Online sustainability information in local governments in an austerity context: an empirical analysis in Italy and Spain", *Online Information Review*, Vol. 40 No. 4, pp. 497-514.

Caba-Pérez, C., Rodríguez-Bolívar, M.P. and López-Hernández, A.M. (2014), "The determinants of government financial reports online", *Transylvanian Review of Administrative Sciences*, Vol. 42, pp. 5-32, available at: www.rtsa.ro/tras/index.php/tras/article/viewFile/15/13

Cegarra-Navarro, J.G., Pachón, J.R.C. and Cegarra, J.L.M. (2012), "E-government and citizen's engagement with local affairs through e-websites: the case of Spanish municipalities", *International Journal of Information Management*, Vol. 32 No. 5, pp. 469-478.

Chandra, A., Suaib, M. and Beg, R. (2015), "Google search algorithm updates against web spam", *Informatics Engineering International Journal*, Vol. 3 No. 1, pp. 1-10.

Collins, B.R. and Flick, E. (1995), "Infofilter: making sense of the internet", *New Review of Information Networking*, Vol. 1 No. 1, pp. 203-207.

Constantinides, E. (2002), "The 4S web-marketing mix model", *Electronic Commerce Research and Applications*, Vol. 1 No. 1, pp. 57-76.

Cuadrado-Ballesteros, B., Frías-Aceituno, J. and Martínez-Ferrero, J. (2014), "The role of media pressure on the disclosure of sustainability information by local governments", *Online Information Review*, Vol. 38 No. 1, pp. 114-135.

da Cruz, N.F., Tavares, A.F., Marques, R.C., Jorge, S. and de Sousa, L. (2016), "Measuring local government transparency", *Public Management Review*, Vol. 18 No. 6, pp. 866-893.

De Andrés, J., Lorca, P. and Martínez, A.B. (2010), "Factors influencing web accessibility of big listed firms: an international study", *Online Information Review*, Vol. 34 No. 1, pp. 75-97.

Ebbers, W.E., Jansen, M.G. and van Deursen, A.J. (2016), "Impact of the digital divide on e-government: expanding from channel choice to channel usage", *Government Information Quarterly*, Vol. 33 No. 4, pp. 685-692.

Einstein, K.L. and Kogan, V. (2016), "Pushing the city limits: policy responsiveness in municipal government", *Urban Affairs Review*, Vol. 52 No. 1, pp. 3-32.

Eurostat (2017a), "Individuals using the Internet for interaction with public authorities", available at: http://ec.europa.eu/eurostat/web/digital-economy-and-society/data/main-tables (accessed 11 January 2018).

Eurostat (2017b), "Statistics on rural areas in the EU", available at: http://ec.europa.eu/eurostat/statistics-explained/index.php/Statistics_on_rural_areas_in_the_EU (accessed 4 April 2018).

Evans, M.P. (2007), "Analysing Google rankings through search engine optimization data", *Internet Research*, Vol. 17 No. 1, pp. 21-37.

Feeney, M.K. and Brown, A. (2017), "Are small cities online? Content, ranking, and variation of US municipal websites", *Government Information Quarterly*, Vol. 34 No. 1, pp. 62-74.

Ferraz Esteves de Araujo, J.F. and Tejedo-Romero, F. (2016), "Local government transparency index: determinants of municipalities' rankings", *International Journal of Public Sector Management*, Vol. 29 No. 4, pp. 327-347.

Ferraz, R. (2015), "Exploring web attributes related to image accessibility and their impact on search engine indexing", *Procedia Computer Science*, Vol. 67, pp. 171-184, available at: www.sciencedirect.com/science/article/pii/S1877050915031075

Fietkiewicz, K.J., Mainka, A. and Stock, W.G. (2017), "eGovernment in cities of the knowledge society. An empirical investigation of smart cities' governmental websites", *Government Information Quarterly*, Vol. 34 No. 1, pp. 75-83.

Gandía, J.L. and Archidona, M.C. (2008), "Determinants of web site information by Spanish city councils", *Online Information Review*, Vol. 32 No. 1, pp. 35-57.

Gandía, J.L., Marrahí, L. and Huguet, D. (2016), "Digital transparency and Web 2.0 in Spanish city councils", *Government Information Quarterly*, Vol. 33 No. 1, pp. 28-39.

García-Sánchez, I.M., Frías-Aceituno, J.V. and Rodríguez-Domínguez, L. (2013), "Determinants of corporate social disclosure in Spanish local governments", *Journal of Cleaner Production*, Vol. 39, pp. 60-72, available at: www.sciencedirect.com/science/article/pii/S0959652612004593

García-Tabuyo, M., Saez-Martin, A. and Caba-Perez, C. (2017), "Proactive disclosure of public information: legislative choice worldwide", *Online Information Review*, Vol. 41 No. 3, pp. 354-377.

Gauld, R., Goldfinch, S. and Horsburgh, S. (2010), "Do they want it? Do they use it? The 'demand-side' of e-Government in Australia and New Zealand", *Government Information Quarterly*, Vol. 27 No. 2, pp. 177-186.

Gelman, A. and Hill, J. (2007), *Data Analysis using Regression and Multilevel Hierarchical Models*, Cambridge University Press, Cambridge.

Hedden, H. (2008), "How semantic tagging increases findability", *EContent*, Vol. 31 No. 8, pp. 1-6.

Huang, Z. and Benyoucef, M. (2014), "Usability and credibility of e-government websites", *Government Information Quarterly*, Vol. 31 No. 4, pp. 584-595.

Huberty, C.J. (1989), "Problems with stepwise methods—better alternatives", in Thompson, B. (Ed.), *Advances in Social Science Methodology*, Vol. 1, Jai Press, Bingley, pp. 43-70.

Ingwersen, P. (1998), "The calculation of web impact factors", *Journal of Documentation*, Vol. 54 No. 2, pp. 236-243.

King, B.A. and Youngblood, N.E. (2016), "E-government in Alabama: an analysis of county voting and election website content, usability, accessibility, and mobile readiness", *Government Information Quarterly*, Vol. 33 No. 4, pp. 715-726.

Kliman-Silver, C., Hannak, A., Lazer, D., Wilson, C. and Mislove, A. (2015), "Location, location, location: the impact of geolocation on web search personalization", *Proceedings of the 2015 Internet Measurement Conference (IMC '15), ACM, New York, NY*.

Kopackova, H., Michalek, K. and Cejna, K. (2010), "Accessibility and findability of local e-government websites in the Czech Republic", *Universal Access in the Information Society*, Vol. 9 No. 1, pp. 51-61.

Lee, J. (2010), "10 year retrospect on stage models of e-Government: a qualitative meta-synthesis", *Government Information Quarterly*, Vol. 27 No. 3, pp. 220-230.

Luh, C.J., Yang, S.A. and Huang, T.L.D. (2016), "Estimating Google's search engine ranking function from a search engine optimization perspective", *Online Information Review*, Vol. 40 No. 2, pp. 239-255.

McNeal, R.S., Tolbert, C.J., Mossberger, K. and Dotterweich, L.J. (2003), "Innovating in digital government in the American states", *Social Science Quarterly*, Vol. 84 No. 1, pp. 52-70.

Margetts, H.Z. (2009), "The Internet and public policy", *Policy & Internet*, Vol. 1 No. 1, pp. 1-21.

Mavridis, T. and Symeonidis, A.L. (2015), "Identifying valid search engine ranking factors in a Web 2.0 and Web 3.0 context for building efficient SEO mechanisms", *Engineering Applications of Artificial Intelligence*, Vol. 41 No. C, pp. 75-91.

Miranda, F.J., Sanguino, R. and Bañegil, T.M. (2009), "Quantitative assessment of European municipal websites: development and use of an evaluation tool", *Internet Research*, Vol. 19 No. 4, pp. 425-441.

Mistry, P., Mistry, D. and Sheth, J. (2013), "Internet marketing: comparative analysis of search engine optimization applications on various parameters", *National Journal of System and Information Technology*, Vol. 6 No. 1, pp. 79-90.

Moreno, L. and Martinez, P. (2013), "Overlapping factors in search engine optimization and web accessibility", *Online Information Review*, Vol. 37 No. 4, pp. 564-580.

Muñoz-Soro, J.F., Esteban, G., Corcho, O. and Serón, F. (2016), "PPROC, an ontology for transparency in public procurement", *Semantic Web*, Vol. 7 No. 3, pp. 295-309.

Ortega, J.L., Orduña-Malea, E. and F. Aguillo, I. (2014), "Are web mentions accurate substitutes for inlinks for Spanish universities?", *Online Information Review*, Vol. 38 No. 1, pp. 59-77.

Page, L. and Brin, S. (1998), "The anatomy of a large-scale hypertextual web search engine", *Proceedings of the Seventh International Web Conference (WWW 98)*, pp. 1-25.

Peristeras, V., Tarabanis, K. and Goudos, S.K. (2009), "Model-driven eGovernment interoperability: a review of the state of the art", *Computer Standards & Interfaces*, Vol. 31 No. 4, pp. 613-628.

Schwarz, G. (1978), "Estimating the dimension of a model", *The Annals of Statistics*, Vol. 6 No. 2, pp. 461-464.

Serrano-Cinca, C., Rueda-Tomás, M. and Portillo-Tarragona, P. (2009), "Determinants of e-government extension", *Online Information Review*, Vol. 33 No. 3, pp. 476-498.

Su, A.J., Hu, Y.C., Kuzmanovic, A. and Koh, C.K. (2014), "How to improve your search engine ranking: myths and reality", *ACM Transactions on the Web (TWEB)*, Vol. 8 No. 2, p. 8.

Tavares, A.F. and Da Cruz, N.F. (2017), "Explaining the transparency of local government websites through a political market framework", *Government Information Quarterly*, in press, available at: https://doi.org/10.1016/j.giq.2017.08.005

Tibshirani, R. (1996), "Regression shrinkage and selection via the Lasso", *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58 No. 1, pp. 267-288.

Tirado-Valencia, P., Rodero-Cosano, M.L., Ruiz-Lozano, M. and Rios-Berjillos, A. (2016), "Online sustainability information in European local governments: an explicative model to improve transparency", *Online Information Review*, Vol. 40 No. 3, pp. 400-415.

Vaughan, L. (2014), "Discovering business information from search engine query data", *Online Information Review*, Vol. 38 No. 4, pp. 562-574.

White, B. (2003), "Web accessibility, mobility and findability", *Proceedings of the IEEE First Latin American Web Congress, LA-WEB 2003*, *Santiago de Chile*, *10-12 November*, pp. 239-241.

Wilson, R.F. and Pettijohn, J.B. (2008), "Using keyword research software to assist in the search for high-demand, low-supply online niches: an overview", *Journal of Internet Commerce*, Vol. 6 No. 1, pp. 101-117.

Wylie, J. (2012), *Make your Website Sell: The Ultimate Guide to Increasing your Online Profits*, Marshall Cavendish Business, London.

Youngblood, N.E. and Mackiewicz, J. (2012), "A usability analysis of municipal government website home pages in Alabama", *Government Information Quarterly*, Vol. 29 No. 4, pp. 582-588.

Zhang, S. and Cabage, N. (2017), "Search engine optimization: comparison of link building and social sharing", *Journal of Computer Information Systems*, Vol. 57 No. 2, pp. 148-159.

Corresponding author
Carlos Serrano-Cinca can be contacted at: serrano@unizar.es

A bibliometric analysis of event detection in social media

Xieling Chen
Jinan University, Guangzhou, China
Shan Wang
University of Macau, Macao, China, and
Yong Tang and Tianyong Hao
South China Normal University, Guangzhou, China

Received 5 March 2018
Revised 26 June 2018
Accepted 11 September 2018

Abstract

Purpose – The purpose of this paper is to explore the research status and development trend of the field of event detection in social media (ED in SM) through a bibliometric analysis of academic publications.
Design/methodology/approach – First, publication distributions are analyzed including the trends of publications and citations, subject distribution, predominant journals, affiliations, authors, etc. Second, an indicator of collaboration degree is used to measure scientific connective relations from different perspectives. A network analysis method is then applied to reveal scientific collaboration relations. Furthermore, based on keyword co-occurrence analysis, major research themes and their evolutions throughout time span are discovered. Finally, a network analysis method is applied to visualize the analysis results.
Findings – The area of ED in SM has received increasing attention and interest in academia with Computer Science and Engineering as two major research subjects. The USA and China contribute the most to the area development. Affiliations and authors tend to collaborate more with those within the same country. Among the 14 identified research themes, newly emerged themes such as Pharmacovigilance event detection are discovered.
Originality/value – This study is the first to comprehensively illustrate the research status of ED in SM by conducting a bibliometric analysis. Up-to-date findings are reported, which can help relevant researchers understand the research trend, seek scientific collaborators and optimize research topic choices.

Keywords Network analysis, Social media, Bibliometric analysis, Event detection

Paper type Research paper

Introduction

Social media are forms of electronic communication (such as websites for social networking and microblogging) through which users create online communities to share information, ideas, personal messages and other contents, such as videos (www.merriam-webster.com/dictionary/social%20media). The ubiquity of social media is increasing rapidly due to the spread of internet and the development of mobile devices. With the continuous growth of social networks and the active use of social media services, overwhelming amount of user-generated social media data is available as participants continuously interact with each other (Weiler *et al.*, 2016). The explosive increasing of the data has promoted the development of social media data mining for supporting better decision making (Wang *et al.*, 2017), thus it draws much attention from both academia and industry (Jiang *et al.*, 2017). Meanwhile, social media data processing brings a lot of research opportunities (Nurwidyantoro and Winarko, 2013), among which, event detection has been one of the most important research topics (Dong *et al.*, 2015) since social media data carry abundant hidden occurrences of real-time events (Kaleel and Abhari, 2015).

Event detection is defined as identifying the first story on the topics of interest through constantly monitoring news streams (Dou *et al.*, 2012). The streaming nature of social media with no limitation about space and time makes it a valuable source for event detection (Nurwidyantoro and Winarko, 2013). Specially, an event in social media has been defined by

This paper forms part of a special section "Social media mining for journalism".

Online Information Review
Vol. 43 No. 1, 2019
pp. 29-52

1468-4527
DOI 10.1108/OIR-03-2018-0068

Dou *et al.* (2012) as "an occurrence causing change in the volume of text data that discusses the associated topic at a specific time. This occurrence is characterized by topic and time, and often associated with entities such as people and location." Thus, event detection in social media (ED in SM) can be defined as the detection of an occurrence causing change in the volume of social media data that discusses the associated topic at a specific time. ED in SM is challenging because of short and noisy contents, diverse and fast changing topics, as well as large data volumes (Petrović *et al.*, 2010; Li *et al.*, 2012).

The research of ED in SM has attracted more and more interests from academia. Researchers have published a wealth of relevant publications to report their research findings. The growing number of publications witnesses its rapid development (e.g. Wang *et al.*, 2017; Hua *et al.*, 2016; Kaleel and Abhari, 2015; Cheng and Wicks, 2014). These research publications, conveying newly developed technologies, reflect the status and trend of the research field to a certain extent. Therefore, a systematic and comprehensive analysis of publications in this research field to help people understand its current status and development trend is of great demand.

However, to the best of our knowledge, currently there is no scientific an
d comprehensive analysis of ED in SM research field based on quantitative and statistical perspective. Therefore, we propose a bibliometric analysis of ED in SM to comprehensively map the landscape of this research field. Specifically, this paper addresses the following three research questions:

RQ1. What are the publication distributions (publication and citation quantities, research subject distribution, predominant journals, countries/regions, affiliations and authors) of the ED in SM research field?

RQ2. What are the scientific collaborations among countries/regions, affiliations and authors?

RQ3. What are the major research themes and their evolutions upon time in this research field?

To answer these questions, this study conducts a bibliometric analysis on academic publications in this field during the period 2009–2017 to explore the general publication distributions, reveal the collaboration relations, discover hot keywords and analyze major research themes and their evolutions.

Literature review

Bibliometric analysis is the evaluation of scholarly publications from a quantitative perspective within a certain field using statistical methods (Chiu and Ho, 2007). It enables researchers to organize information in a specific field (Merigo *et al.*, 2015), evaluate scientific developments in the knowledge of a specific subject (Bouyssou and Marchant, 2011), compare research performance across different countries (Pu *et al.*, 2016), identify emerging research focuses and predict future research success (Mazloumian, 2012), etc. Bibliometric analysis has been widely applied to various fields for the measurement of quality and productivity of academic outputs, e.g., public service management (Juliani and de Oliveira, 2016), multimorbidity (Xu *et al.*, 2017) and fuzzy theory (Yu *et al.*, 2018). Especially, it has also been applied to interdisciplinary research fields, e.g., natural language processing in mobile computing (Chen, Ding, Xu, Wang, Hao and Zhou, 2018), and natural language processing in medical research (Chen, Xie, Wang, Liu, Xu and Hao, 2018). Specially, some bibliometric studies centered on social media-related topics. Leung *et al.* (2017) provided a systematic review on 406 social media-related literatures during the period 2007–2016 with co-citation and co-word analysis. Three topics including big data using text mining methods, diverse research designs and word-of-mouth based on social theories were

identified as future research directions. Zyoud *et al.* (2018) assessed the publication growth, international collaboration, author productivity and emerging topics based on literatures related to social media in the field of psychology. Li, Wei, Xiong, Feng, Ye and Jiang (2017) conducted a bibliometric analysis to review social media research during the period 2008–2014 based on publications indexed in science citation index and social science citation index databases.

Various techniques involve in the process of ED in SM. For identifying irrelevant or unverified tweets, different classification algorithms, e.g., support vector machine (SVM), logistic regression, multilayer perceptron, Naive Bayes, etc., can be used. SVM and perceptron outperform other comparative techniques especially for opining detection toward certain event (Bravo-Marquez *et al.*, 2014). Event detection techniques for opinion mining is conducted based on locality sensitive hashing, intermediate semantic entity, dominant entity, entity linking, etc. Social data clustering is frequently adopted for the research of ED in SM, in which techniques, such as *k*-means algorithm, hypergraph-based algorithm (Amato *et al.*, 2018) and network-based modeling (Gao and Liu, 2017), are widely applied. In addition, named entity recognition is another technique to identify essential entities in the process of event detection. Topic modeling is another major research technique for ED in SM, in which methods such as latent semantic analysis, probabilistic latent semantic analysis and latent Dirichlet allocation are commonly used for topic modeling. As one of the recent work on topic modeling, a non-parametric Bayesian model hierarchical Dirichlet processes was used in sub-story detection in social media (Srijith *et al.*, 2017).

Similar to event detection, opinion mining also aims to summarize core information for a volume of text data. Compared with event detection that focuses more on issue description, opinion mining concerns more about public opinions toward certain topics. Taraghi *et al.* (2013) addressed problem about the recommendation of appropriate relevant work from an increasing amount of scientific literatures. They introduced a recommender system, in which the analysis of user paths, i.e. the sequence of articles read by users, was a main focus. Petz *et al.* (2015) investigated the differences among various social media channels with respect to opinion mining. They further conducted evaluation on the effectiveness of various text preprocessing algorithms in these channels as a subtask of opinion mining. Petz *et al.* (2012) discussed preprocessing techniques for dealing with emerging problems occurring in opining mining for real world situations.

As for the research field of ED in SM, some relevant systematic reviews can be found. Dong *et al.* (2015) explored and summarized popular tasks in the social media event detection field. Nurwidyantoro and Winarko (2013) described topics related to the analysis of social media for detecting event types of disaster, traffic, outbreak and news. Atefeh and Khreich (2015) provided a survey of techniques for event detection from Twitter streams.

Research methods

Data collection

Web of Science (WoS) is the most authoritative academic publication and citation repository, especially the core collection database of WoS composed of "Science citation index expanded," "Social sciences citation index," and "Arts and humanities citation index." Therefore, they were used as the data source to retrieve relevant publications. Publications were identified using "TS" (Topics) as a search field, referring to the title, abstract or keywords of a publication. With a well-defined query (shown below "query for retrieving research publications from web of science"), 673 publications in "Article" and "Proceedings paper" types during the period 2009–2017 were retrieved on February 1, 2018. All the retrieved publications were downloaded as plain texts. In order to ensure that they were

closely related to the research field, manual verification was conducted by a domain expert through manually reviewing each publication based on the criteria listed in Table I.

The query for retrieving research publications from web of science:

- TS = (("event") AND ("social media" OR "social medium" OR "online media" OR "online news" OR "online information" OR "online text" OR "online sites" OR "webpages" OR "Web news" OR "Web information" OR "Web media" OR Twitter OR Twit OR "online community" OR "Facebook" OR "WeChat" OR "Weibo" OR "YouTube" OR "LinkedIn" OR "Instagram" OR "online blogs" OR "microblogs" OR "WhatsApp" OR "virtual community" OR "social networking site" OR "social bookmarking" OR "consumer-generated media")).

In total, 565 publications were finally selected for analysis. Based on them, key elements including title, author, journal, publication date, subject category, language, funding, author keywords, keywords-plus, abstract, author address, as well as citation count, page count and reference count were extracted. In addition, corresponding affiliations and countries/regions were extracted from author address information. Table II shows the statistics of the elements. The average citation count and reference count of the publications are 7.46 and 42.57, respectively.

Data analysis strategies

Bibliometric analysis mainly centers on the exploration of publication distributions, collaboration relations, as well as research topic distributions and evolutions. In this study, a distribution analysis is applied to discover the general publication distributions in the research area of ED in SM, i.e. publication and citation quantities, research subject

Table I.
The inclusion and exclusion criteria for manually verifying the retrieved publications

Inclusion criteria	
I1	Information quality, trust and credibility aspects in the field
I2	Novel techniques, methods and strategies in the field
I3	Social, political and ethical aspects in the field
I4	Big data processing technology aspect in the field
Exclusion criteria	
E1	Irrelevant to event detection
E2	No abstract

Table II.
The statistics of the key elements of the publications

Characteristics	Statistics	Characteristics	Statistics
Total number of publications	565	Average number of affiliations in one publications	2.09
Number of unique journals	310	Average number of authors in one publications	3.74
Number of unique countries (or regions)	52	Average number of citations for one publications	7.46
Number of unique affiliations/first affiliations	726; 428	Average number of references in one publications	42.57
Number of unique authors/first authors/last authors	1,866; 542; 526	Average number of author keywords or keywords-plus	7.99
Number of publications with single country/affiliation/author	402; 221; 49	Average number of words/characters in title	11.89; 84.75
Average number of countries (or regions) in one publications	1.36	Average number of words/characters in abstract	197.24; 1,337.10

distribution, as well as predominant journals, countries/regions, affiliations and authors, to answer *RQ1*. Especially, in the analysis of publication and citation quantities, a regression modeling is also used to approximate the development trends of publication and citation quantities with time. To answer *RQ2*, an indicator of collaboration degree is applied to measure research's connective relations in the perspectives of countries/regions, affiliations and authors. A network analysis method is further conducted to reveal the scientific collaboration relations. As for *RQ3*, this study identifies main research themes through clustering analysis based on keyword co-occurrence analysis. Following that, the emergence and evolution of keywords are mined through the comparison of two periods, i.e. 2009–2013 and 2014–2017. The keyword analysis and clustering analysis results are further visualized using a network analysis method. Python and R programming languages are used to perform network analysis, keyword co-occurrence analysis and clustering analysis.

Findings of publication distribution analysis

The general publication distributions of the research of ED in SM are presented in this section, including the trends of publication and citation quantities, subject distribution, as well as predominant journals, authors, affiliations and countries.

Publication and citation quantities

Through the statistical calculation on the retrieved data, the distributions of total publications and average citations by year are presented in Figure 1. We conduct liner, logarithmic, polynomial, exponential and power curves regression analysis with year as independent variable x for fitting the trends of total publications and average citations. In the regression modeling, period 2009–2016 is used for regression fitting, while year 2017 is used to validate the predictive effect of the fitting curve. Two polynomial regression curves with the highest R^2 are selected out as the optimal models, which are also integrated in the figure.

From the result, there is an increasing trend of the publication quantity, reflecting a growing research enthusiasm in the field. This is also demonstrated by the positive coefficient of χ^2 in the estimated regression model ($R^2 = 0.9827$). The predictive value of the estimated regression model for year 2017 is calculated as: $2.738096 \times 2{,}017^2 - 11{,}002.643 \times 2{,}017 + 11{,}053{,}140 = 174.91$, which is very close to the actual value 173, and the actual value also falls within a 95% confidence interval as (139.69, 204.17).

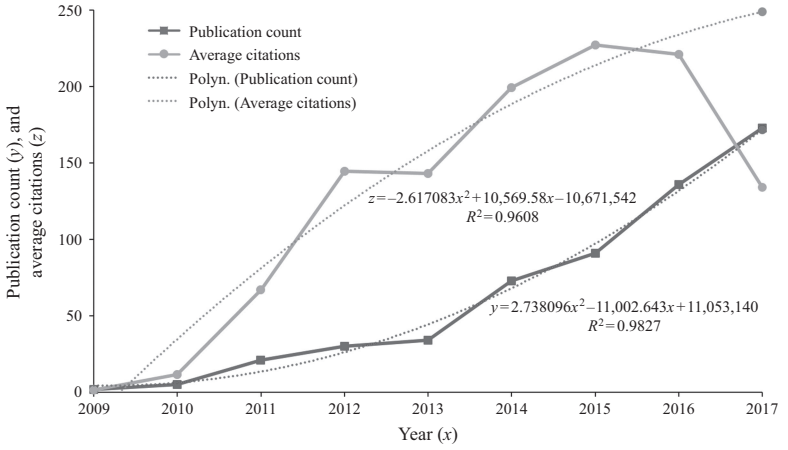


Figure 1.
Total publications and average citations

To alleviate the duration influence, we use average citations instead of total citations since total citations are increasing with time and there is no annual citation information being provided yet. The calculation of average citations for a publication p_i is as Equation (1). The average citations present an upward trend until 2015, showing an increasing influence of the research. The estimated regression model for average citations does not demonstrate a prediction effect as good as the estimated regression model for publication quantity since new publications usually have less citations due to the limited time:

$$\text{Average citations}(p_i) = \text{Total citations}(p_i)/(2017-\text{publishing year}(p_i)+1) \quad (1)$$

Research subject distribution

Research subject, denoting the research areas, is an important element for measuring the performance of a research field based on publication and citation percentages (Ab Razak *et al.*, 2016). In the analysis of subject distribution, the WoS subject category is utilized, in which 72 subjects in total are identified to be relevant to our analysis. Due to the interdisciplinary nature of some research, a publication may belong to more than one subject. For the 565 publications, Figure 2 shows the top 10 subjects by the quantity of publications and their citations, respectively.

The result illustrates that Computer Science is the most popular category, taking up nearly 30 percent of the total publications and 32 percent of the total citations. Engineering is ranked at 2nd accounting for 8.07 percent of the total publications and 8.16 percent of the total citations. The two subjects together contribute a lot to the development of the research field. For most of the ten subjects, the publication percentage and citation percentage are relatively close. It is worth noticing that Communication and Telecommunication possess comparatively higher citation percentages than their publication percentages. This reflects the high influence and quality of the publications in the two subjects.

Predominant journals

This section reveals predominant journals in terms of publication and citation quantities. This study identifies that 310 journals have published relevant research work, along with their journal impact factors (IF), JCR quartile in category (*Q*), and *H*-index. IF is created by the Institute of Scientific Information according to one journal's citations and publications within a given subject category (Rey-Marti *et al.*, 2016). *Q* shows the relative position of a journal along the range of an impact factor distribution. *H*-index quantifies one's research output as a single figure, meaning that *h* of one's total publications are cited at least *h* times (Hirsch, 2005). It is an objective indicator considering both quantity and quality of one's academic level (Alonso *et al.*, 2009). In calculating *H*-index of journals, we rank all the *n* publications of each journal *j* from 1 to *n* by citation counted to February 1, 2018 in descending order. The No. *h* publication has a citation no less than *h*, while the No. (*h+1*) publication has a citation less than (*h+1*), then *h* is the *H*-index for the journal *j*. Table III presents the top

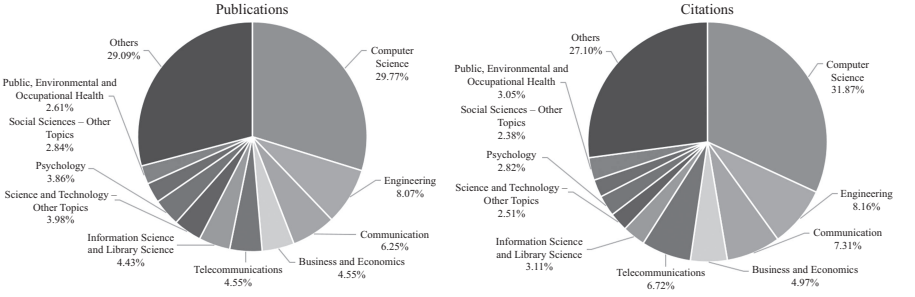


Figure 2.
Publication and citation distribution of the top 10 subjects

Table III.
Top journals with the most publications and citations

Most productive journals	TP↓	TC	ACP	*H*	IF (*Q*)	Most cited journals	TC↓	TP	ACP	*H*	IF (*Q*)
PLoS One	26	149	5.73	7	2.806 (Q1)	*IEEE Transactions on Multimedia*	245	9	27.22	7	3.509 (Q1)
Multimedia Tools and Applications	16	74	4.63	5	1.53 (Q2)	*IEEE Transactions on Visualization and Computer Graphics*	223	7	31.86	6	2.84 (Q1)
Computers in Human Behavior	14	54	3.86	4	3.435 (Q1)	*Information Communication & Society*	186	8	23.25	5	2.692 (Q1)
IEEE Transactions on Multimedia	9	245	27.22	7	3.509 (Q1)	*International Journal of Computer Vision*	151	2	75.50	2	8.222 (Q1)
Information Communication & Society	8	186	23.25	5	2.692 (Q1)	*PLoS One*	149	26	5.73	7	2.806 (Q1)
Neurocomputing	8	37	4.63	4	3.317 (Q1)	*Drug Safety*	138	6	23.00	4	3.435 (Q1)
World Wide Web-internet and Web Information Systems	8	112	14.00	4	1.405 (Q3)	*World Wide Web-internet and Web Information Systems*	112	8	14.00	4	1.405 (Q3)
Expert Systems with Applications	7	97	13.86	3	3.928 (Q1)	*Internet Research*	111	5	22.20	4	2.931 (Q1)
IEEE Transactions on Knowledge and Data Engineering	7	94	13.43	2	3.438 (Q1)	*Transactions in GIS*	105	3	35.00	2	2.252 (Q2)
IEEE Transactions on Visualization and Computer Graphics	7	223	31.86	6	2.84 (Q1)	*Government Information Quarterly*	101	2	50.50	1	4.09 (Q1)

Notes: TP, total publications; TC, total citations; ACP, average citations per publication, calculated as TC/TP; *H*, *H*-index; IF (*Q*), impact factor and JCR quartile in category for year 2016

10 journals with the most publications, as well as the top 10 journals with the most citations. The top 10 most productive journals in the left column only account for 19.47 percent of the total publications. This reflects a diversified distribution of these publications and a broad interest from multiple research perspectives. The top 3 productive journals are *PLoS One*, *Multimedia Tools and Applications*, and *Computers in Human Behavior*. *PLoS One*, as a multidisciplinary journal with broad scopes, also has the highest *H*-index. Regarding the most cited journals, *IEEE Transactions on Multimedia* receives the most citations as well as the highest *H*-index over the years, followed by *IEEE Transactions on Visualization and Computer Graphics*. The former focuses on multimedia technology and applications, while the latter covers the breadth of research related to computer graphics and visualization techniques. It is worth noticing that *International Journal of Computer Vision* possesses the highest ACP as 75.5, reflecting high quality of its publications.

Predominant countries/regions

All the affiliation countries/regions, affiliations and authors participating in each publication are used for the analysis of predominant ones. The 565 publications originate from 52 countries/regions. Table IV presents the statistics of the top 10 countries/regions with regard to the total publications in this field. The USA is the most productive country with 206 publications (36.46 percent), followed by China with 111 publications (19.65 percent). The two countries together account for 56 percent of the total publications, indicating their huge enthusiasm in the research field. Furthermore, the USA possesses the highest *H*-index of 21, showing that it not only dominates in the publication count, but also has a high academic influence in the research field. China is the only developing country among the ten countries. Its *H*-index is ranked at 2nd ($H = 15$), indicating its high productivity. Overall, the ACP of internationally collaborated publications for most countries/regions is much higher than those without international collaboration. This indicates that international collaboration can potentially improve the quality of publications. The USA and China collaborate with each other for 19 times, reflecting their close collaboration in the research field. Another interesting finding is that Japan is the one with the least international collaboration (16.67 percent) among the ten countries, suggesting the potential to encourage their scholars to engage in international collaboration.

						Single-country/region		International collaboration	
Rank	Country	TP (%)	TC	ACP	*H* (*R*)	ACP	*P*%	ACP	TFC (*n*)
1	USA	206 (36.46)	1,857	9.01	21 (1)	8.57	64.56	9.82	China (19)
2	China	111 (19.65)	759	6.84	15 (2)	5.10	52.25	8.74	USA (19)
3	England	59 (10.44)	547	9.27	12 (3)	7.28	49.15	11.20	USA (8)
4	Australia	50 (8.85)	491	9.82	11 (4)	7.36	56.00	12.95	China (10)
5	Germany	32 (5.66)	184	5.75	7 (6)	5.76	65.63	5.73	England (4)
6	Italy	29 (5.13)	124	4.28	6 (8)	3.85	44.83	4.63	USA (5)
7	South Korea	22 (3.89)	127	5.77	6 (8)	2.42	54.55	9.80	USA (5)
8	Canada	21 (3.72)	157	7.48	6 (8)	6.57	33.33	7.93	USA (5)
9	Japan	18 (3.19)	207	11.50	5 (12)	13.67	83.33	0.67	USA (2)
9	Spain	18 (3.19)	116	6.44	3 (17)	2.92	72.22	15.60	USA (2)

Table IV.
Top productive countries/regions ranked by total publications

Notes: TP (%), publication count and percentage; TC, total citations; ACP, average citations per publication; *H* (*R*), *H*-index of a country/region and its rank; *P*%, percentage of single-country/region publications within its total publications; TFC (*n*), the closest collaborator and collaboration times

The annual publications of the top 4 countries exhibit an upward trend, as shown in Figure 3. The USA has a leading position during nearly the entire period except the year 2009. The publication counts for the four countries experience sharp increases during the period 2014–2017. In particular, the publication count of the USA for the period 2014–2017 (176 publications) are nearly six times than that for the period 2009–2013 (30 publications), while China has nine times increase.

Productive affiliations

There are 726 affiliations from all over the world having relevant research work among the publications. Table V presents the top 12 productive affiliations. Six of them originate from the USA, which again indicates the USA's dominant position in the research field.

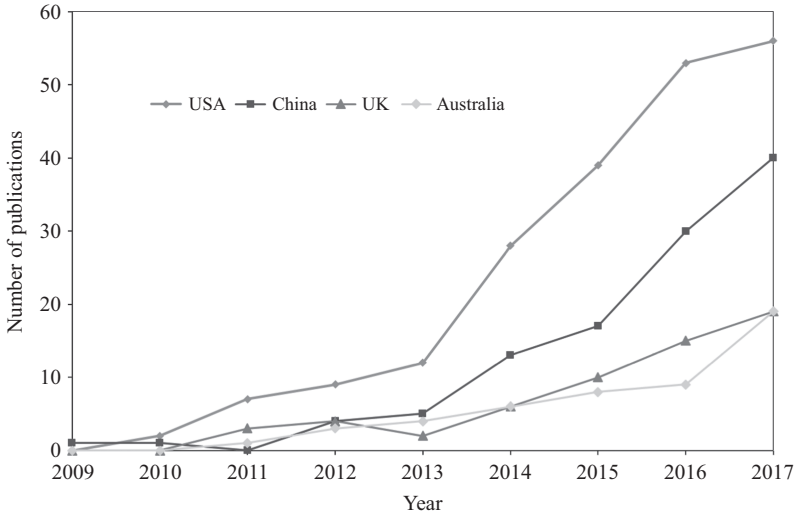


Figure 3.
Annual publication distributions of the top 4 countries

Rank	Name	Country	QS Rank	TP (%)	TC	ACP	*H* (*R*)	FP (*R*)	CP (%)
1	Tsinghua University	China	25	21 (3.72)	221	10.52	7 (1)	15 (1)	18 (85.71)
2	Shanghai University	China	421–430	17 (3.01)	158	9.29	5 (3)	8 (4)	15 (88.24)
3	The Ministry of Public Security	China	n/a	15 (2.65)	170	11.33	6 (2)	10 (2)	15 (100.00)
4	Chinese Academy of Sciences	China	n/a	14 (2.48)	89	6.36	5 (3)	9 (3)	14 (100.00)
5	Virginia Polytechnic Institute and State University	USA	367	11 (1.95)	122	11.09	3 (14)	5 (9)	9 (81.82)
6	University at Albany SUNY	USA	651–700	9 (1.59)	13	1.44	2 (39)	3 (32)	9 (100.00)
7	Arizona State University	USA	209	7 (1.24)	85	12.14	3 (14)	2 (49)	6 (85.71)
7	Beihang University	China	551–600	7 (1.24)	89	12.71	4 (5)	7 (5)	5 (71.43)
7	George Mason University	USA	801–1,000	7 (1.24)	137	19.57	4 (5)	7 (6)	4 (57.14)
7	National University of Singapore	Singapore	15	7 (1.24)	84	12.00	4 (5)	4 (15)	6 (85.71)
7	University of Maryland	USA	129	7 (1.24)	59	8.43	3 (14)	4 (21)	7 (100.00)
7	University of Pennsylvania	USA	19	7 (1.24)	34	4.86	3 (14)	3 (41)	5 (71.43)

Notes: QS Rank, QS World University Rankings 2018; TP (%), publication count and percentage; TC, total citations; ACP, average citations per publication; *H* (*R*), *H*-index of an affiliation and its rank; FP (*R*), publication count as the first affiliation and its rank; CP (%), collaborated publication count and percentage

Table V.
Top affiliations ranked by total publications

Five affiliations including the top 4 are located in China and the most productive one is Tsinghua University with the highest *H*-index. This again confirms that China is one of the active countries on the research. George Mason University has the highest ACP although with only seven publications. This indicates the high quality and influence of its publications. The collaboration rates for most of the affiliations are above 80 percent. Especially, The Ministry of Public Security, Chinese Academy of Sciences, SUNY Albany, and University of Maryland have 100 percent publication collaborations with other affiliations. Beihang University and George Mason University serve as first affiliations in all their publications.

Productive authors

The most productive authors are identified. Table VI shows the top 10 authors ranked by publication count. Seven of them come from China, and seven are from the USA. Specially, all the top 4 authors are from China, demonstrating its active role in the research field. The most productive one is Xu Zheng with both the highest publication count and *H*-index, indicating his strong academic productivity. Three authors from the USA have relatively low *H*-indexes. Considering the first author only, Xu Zheng has the most first author publications.

Findings of collaboration analysis

The distributions of annual collaborated publications in the perspectives of country, affiliation and author are presented in Figure 4. In general, collaborated publication counts for the three perspectives increase over time. The figure also presents the calculated collaboration degrees. The collaboration degree is a measure of scientific research's connective relations to the level of authors, affiliations and countries (Wei *et al.*, 2014), with the following equation:

$$D = \left(\sum_{i=1}^{N} \alpha_i \right) / N \tag{2}$$

In the equation, D represents the collaboration degree (Wei *et al.*, 2013). α_i denotes the author count, affiliation count or country count for a publication. N is the annual publication count in the research field.

Out of all the publications, on average each one has three authors and two affiliations since 2012. The international collaboration and affiliation collaboration are growing

Rank	Name	Country	TP	TC	ACP	*H* (*R*)	FP (*R*)	LP (*R*)
1	Xu, Zheng	China	15	170	11.33	6 (1)	10 (1)	4 (2)
2	Luo, Xiangfeng	China	12	110	9.17	5 (2)	2 (6)	1 (39)
3	Mei, Lin	China	9	137	15.22	5 (2)	0	2 (9)
3	Xuan, Junyu	China	9	80	8.89	4 (4)	2 (6)	1 (39)
5	Ramakrishnan, Naren	USA	8	19	2.38	2 (21)	0	7 (1)
5	Zhang, Hui	China	8	62	7.75	4 (4)	0	0
7	Chen, Feng	USA	7	13	1.86	2 (21)	1 (29)	0
8	Hu, Chuanping	China	6	105	17.50	4 (4)	0	2 (9)
8	Liu, Yunhuai	China	6	89	14.83	4 (4)	0	0
8	Lu, Chang-Tien	USA	6	11	1.83	1 (74)	0	0

Table VI.
Top productive authors ranked by total publications

Notes: TP, publication count; TC, total citations; ACP, average citations per publication; *H* (*R*), *H*-index of an author and its rank; FP (*R*), publication count as the first author and its rank; LP (*R*), publication count as the last author and its rank

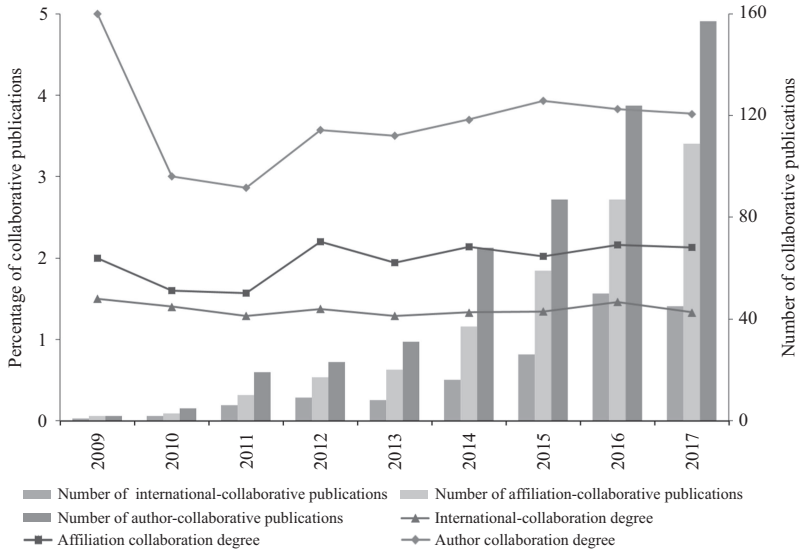


Figure 4.
The distributions of various collaborative publications and collaboration degrees

relatively slowly. This suggests that the authors tend to collaborate more with those within the same affiliation or country. To investigate such influence by research subjects, we calculate collaboration degrees for the top 10 major research subjects. The results are reported in Figure 5. In the perspectives of author collaboration degree and affiliation collaboration degree, subject *Public, Environmental & Occupational Health* achieves the highest values as 6.78 and 3.09. In the perspective of international collaboration degree, subject Telecommunications has the highest value as 1.70. Furthermore, two ratios indicating the level of international collaboration and across affiliation proportion are defined as Equation (3) and Equation (4), respectively. The greater the ratios are, the more

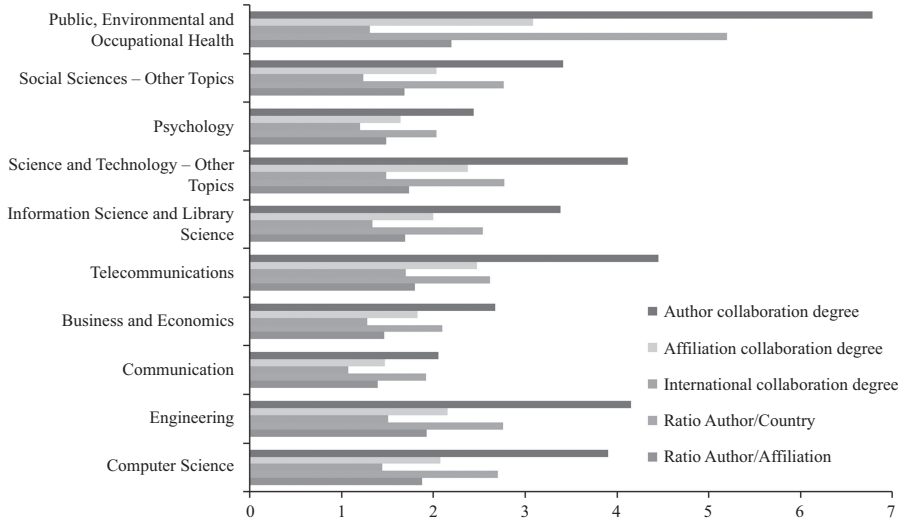


Figure 5.
The distributions of collaboration degree and collaboration ratios by research subjects

the authors tend to collaborate across countries or affiliations:

$$\text{Ratio}\frac{\text{Author}}{\text{Country}} = \frac{\text{Internetional collaboration degree}}{\text{Author collaboration degree}} \tag{3}$$

$$\text{Ratio}\frac{\text{Author}}{\text{Affiliation}} = \frac{\text{Affiliation collaboration degree}}{\text{Author collaboration degree}} \tag{4}$$

Furthermore, the collaboration among countries/regions is visualized as Figure 6 with 52 nodes and 143 links as well as five sparse nodes (Slovakia, Iran, Macau, Egypt and Turkey) using network analysis, which can be accessed via www.zhukun.org/haoty/resources.asp?id=JOIR_cocountry. Each country/region is presented as a node with the node size representing its proportion of publications and the node color denoting the continent that it belongs to. The thickness of each line indicates collaboration strength between two countries/regions. From the figure, the USA (the largest node in blue color) has the most collaborations with other countries/regions. China (the second largest node in orange color) is a leading developing country on its international collaboration. The USA–China collaboration (the thickest line) is ranked at 1st, followed by China–Australia and the USA–Australia collaborations. In addition, the collaboration network between European countries (the nodes in green color) is very dense. The collaboration among affiliations shown as Figure 7 can be accessed via www.zhukun.org/haoty/resources.asp?id=JOIR_coaffliation, and the collaboration among authors shown as Figure 8 can be accessed via www.zhukun.org/haoty/resources.asp?id=JOIR_coauthor. In total, 45 of the 54 affiliations with publications $\geqslant 4$ involve in

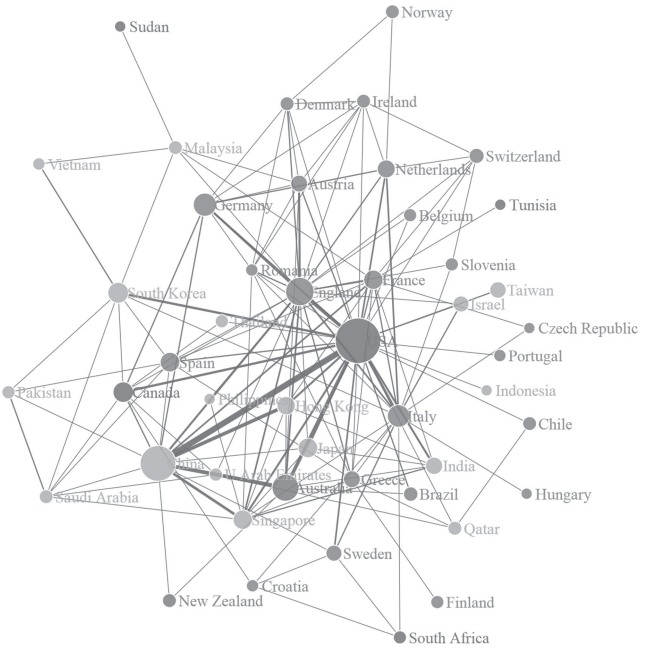
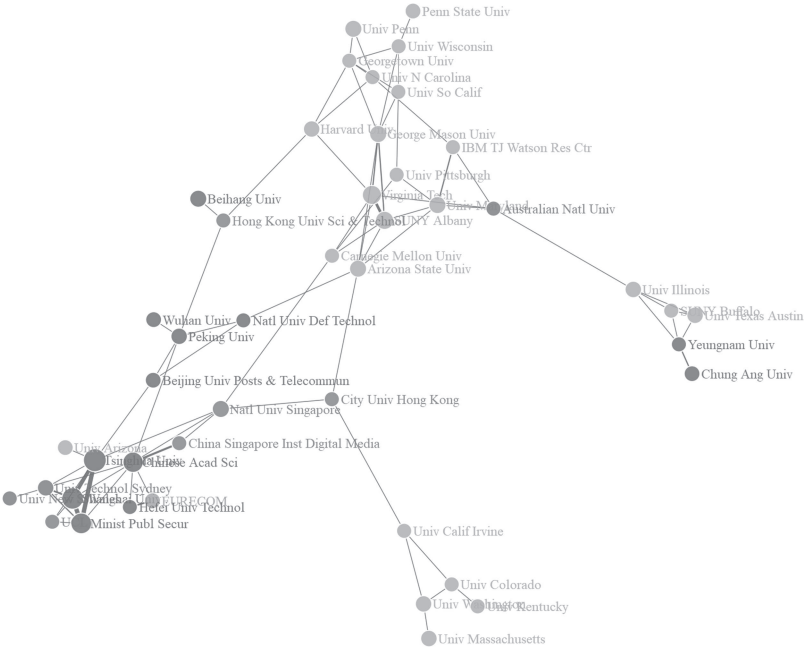


Figure 6.
Collaboration network among countries/regions

Notes: The green nodes represent countries/regions from Europe, orange for Asia, brown for Africa, blue for North America, red for Oceania and purple for South America



Notes: The orange nodes represent affiliations from the USA, blue for China and red for Australia

Figure 7.
Collaboration network of affiliations with publications ⩾4

publication collaboration. In the network, the node count and edge count are 54 and 79 with 9 sparse nodes (Univ Melbourne, Cardiff Univ, Queensland Univ Technol, Drexel Univ, Nanyang Technol Univ, Univ Konstanz, Univ Leeds, Univ Pisa and Univ Tokyo). Among 42 authors with publications ⩾ 3, there exists collaboration among 30 of them. In the network, the node count and edge count are 42 and 59 with 12 sparse nodes (Bruns, Axel, Jung, Jai E., Chen, Hsinchun, Collier, Nigel, Gao, Chao, Lee, Chung-Hong, Poblete, Barbara, Procter, Rob, Sasahara, Kazutoshi, Spence, Patric R., Winter, Stephan, and Yan, Shuicheng). Most of the affiliations come from China and most of the authors come from the USA and China. By accessing to the dynamic networks, users can explore the collaboration relations for specific countries/regions, affiliations or authors by simply clicking the nodes.

Findings of research themes and their evaluations analysis

Keywords co-occurrence analysis

The keywords of a publication to a great extent represent the major research focus of the publication, and, thus, they can help readers quickly identify the research topic of the publication (Zhong *et al.*, 2016). Consequently, keyword co-occurrence can be used for detecting research topics as well as monitoring the transitions of research frontiers in a certain knowledge domain (Lee and Su, 2010; Liu *et al.*, 2015). Based on a keyword co-occurrence analysis, a clustering analysis method is applied to explore the major research hotspots. The emergence and evolution of keywords are identified through the comparison of two periods, i.e., 2009–2013 (92 publications) and 2014–2017 (473 publications). The length of periods is determined in order to acquire a moderate publication count for each period, as well as considering a moderate time span. Furthermore, a network analysis is conducted to visualize the keyword analysis and clustering analysis results.

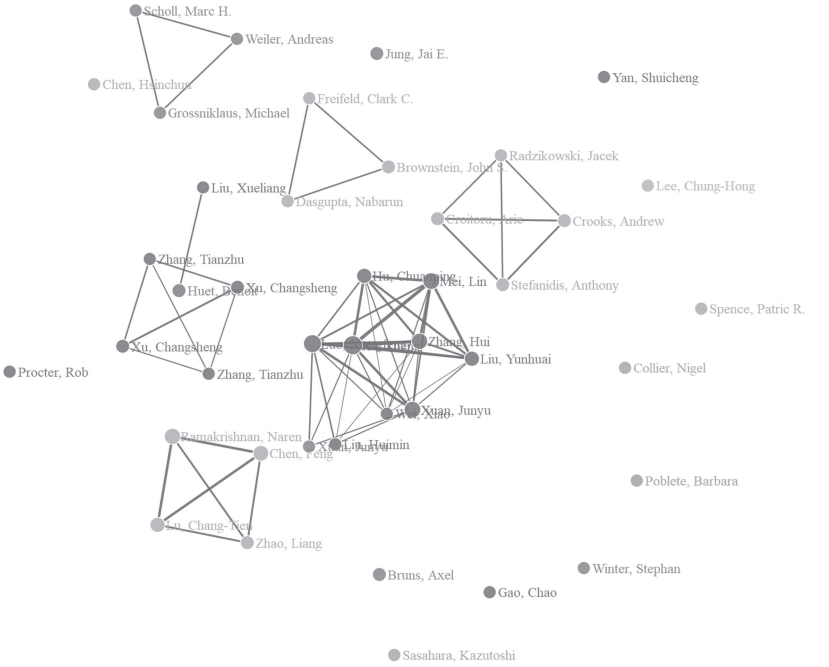


Figure 8.
Collaboration network of authors with publications ⩾3

Notes: The blue nodes represent authors from China, and orange for the USA

In order to improve the effectiveness of keywords analysis, a duplication checking process is conducted (Cobo *et al.*, 2015). Abbreviations are replaced by corresponding full names, e.g., LDA is replaced by latent Dirichlet allocation; NLP is replaced by natural language processing; SVM is replaced by support vector machine. Keywords representing the same concepts are grouped, e.g., behaviour and behavior. Some meaningless keywords such as stop words or words with a very broad and general meaning (i.e. information, system, event, model, algorithm, design, framework, technology, time, tool, issue and people) are removed. Furthermore, to facilitate an efficient analysis, keywords that do not meet a co-occurrence frequency of three are excluded. Overall, 107 keywords meet the threshold after exclusion. The top 20 keywords ranked by frequency are shown in Table VII. The keyword social media is ranked at 1st with a frequency of 179. Other high frequent keywords include

Table VII.
Top 20 keywords ranked by total frequency for the two periods

		Frequency					Frequency		
R	Keywords	Total	2009–2013	2014–2017	R	Keywords	Total	2009–2013	2014–2017
1	Social media	179	25	154	11	Microblog	24	8	16
2	Twitter	113	14	99	12	Communication	23	1	22
3	Event detection	55	11	44	13	Facebook	21	4	17
4	Network	40	7	33	14	Impact	21	1	20
5	social network	32	6	26	15	Text mining	21	7	14
6	Internet	31	11	20	16	Management	20	3	17
7	Big data	26	2	24	17	Disaster	17	0	17
8	Behavior	25	4	21	18	Risk	16	0	16
9	News	25	5	20	19	Sentiment analysis	16	0	16
10	Media	24	2	22	20	Online	14	2	12

Twitter (113), event detection (55) and network (40). Compared with the period 2009–2013, most keywords present a significant growth in frequency during the period 2014–2017. Especially, keywords such as disaster, risk and sentiment analysis have no occurrence during the period 2009–2013 but show significant growth during the period 2014–2017. This again confirms the increasing interest and diversified research focuses in the research of ED in SM.

Major research themes

Based on keyword co-occurrence, a keyword correlation matrix is built and calculated using Ochiai correlation coefficient, which is a measurement of the distance between two keywords. The calculation formula is expressed in the following equation:

$$O_{ij} = A_{ij}/\sqrt{A_i A_j}. \tag{5}$$

In the equation, O_{ij} is the co-occurrence probability of keywords W_i and W_j. A_{ij} indicates the co-occurrence frequency of the two keywords. A_i and A_j are the frequencies of the keywords. With the correlation matrix, hierarchical clustering analysis using complete-linkage is conducted. Figure 9 shows that the 107 frequent keywords are grouped into 14 clusters automatically. By analyzing the semantics of representative terms and reviewing relevant abstracts, potential themes for each cluster is concluded. As shown in Table VIII, we totally identify 14 major research themes including Twitter-related research, Crisis communication research, Pharmacovigilance event detection, Sport related event, Experiment credibility, Text mining research, Health surveillance, Blog community review, etc.

A brief explanation for some themes is presented as follows. Twitter-related research centers on the analysis of Twitter data. Twitter is an important platform in social media research because it provides a large source of publicly available posts on major events. A significant amount of research studying ED in SM collection analyzes data using the Twitter API (e.g. Srijith *et al.*, 2017). Crisis communication research focuses on the use of social media data for the study of crisis communication, which is an essential aspect of disaster and crisis management for governments (Atkinson, 2014). As an open space for the public's opinion and expression, social media has become an increasingly essential issue in crisis events (Luo and Zhai, 2017). Pharmacovigilance event detection comprises research on pharmacovigilance event discovery using social media data. Pharmacovigilance aims to uncover and understand harmful side-effects of drugs, termed adverse events (Yeleswarapu *et al.*, 2014). Experiment credibility is about the evaluation and comparison of different methods for social media mining (e.g. Xuan *et al.*, 2017). Text mining research is the computational process of extracting meaningful information from large amounts of unstructured social media text (Harpaz *et al.*, 2014). Techniques such as natural language processing and event extraction are also included. Health surveillance research refers to health surveillance and communication on events such as influenza (Nawaz *et al.*, 2017) and adverse drug reaction (Yang and Yang, 2015) through the analysis of social media data. Blog community review research is the analysis of blog community review (e.g. Stephen and Galak, 2012). Online emotion detection focuses on emotional disclosure in social media with information technology. Online politic-related event mining includes research relative to the detection of political events such as election campaign (Lorentzen, 2016) from social media data. Network analysis and visualization refers to the techniques of network analysis and visualization for social media research. Visual representation of social networks is important to understand network data and convey the result of the analysis (Serrat, 2017). Emergent event management includes research on the use of social media in emergency management.

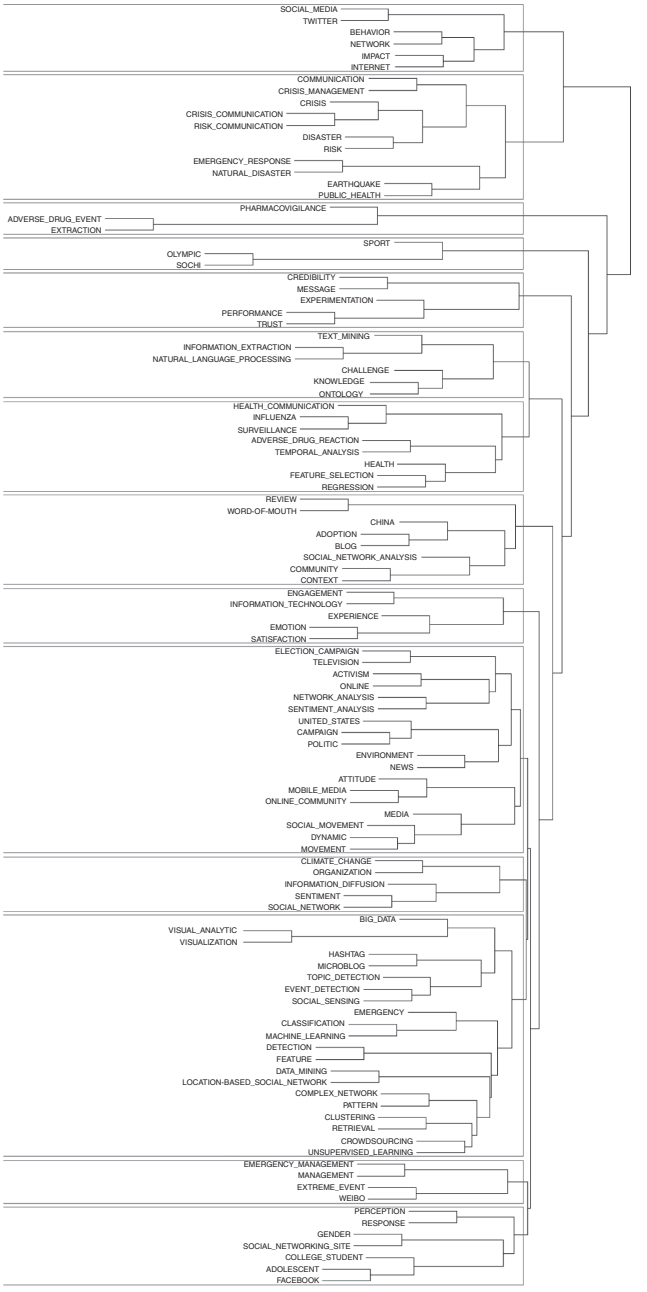


Figure 9.
The visualized result of hierarchical clustering

	Potential themes	2009–2013	2014–2017
1	Twitter-related research	Behavior; impact; internet; network; social media; Twitter	Behavior; impact; internet; network; social media; Twitter
2	Crisis communication research	Communication; earthquake; natural disaster; public health	Communication; *crisis; crisis communication; crisis management; disaster*; earthquake; *emergency response*; natural disaster; public health; *risk; risk communication*
3	Pharmacovigilance event detection		*Adverse drug event; extraction; pharmacovigilance*
4	Sport related event		*Olympic; Sochi; sport*
5	Experiment credibility	Credibility; experimentation; performance; trust	Credibility; experimentation; *message*; performance; trust
6	Text mining research	Challenge; information extraction; knowledge; natural language processing; ontology; text mining	Challenge; information extraction; knowledge; natural language processing; ontology; text mining
7	Health surveillance	Feature selection; health; regression; surveillance; temporal analysis	*Adverse drug reaction*; feature selection; health; *health communication; influenza*; regression; surveillance; temporal analysis
8	Blog community review	Adoption; blog; China; community; context; review; social network analysis; word-of-mouth	Adoption; blog; China; community; context; review; social network analysis; word-of-mouth
9	Online emotion detection	Emotion; engagement; experience; information technology; satisfaction	Emotion; engagement; experience; information technology; satisfaction
10	Online politic-related event mining	Activism; campaign; environment; media; mobile media; news; online; online community; politic; social movement; television; united states	Activism; *attitude*; campaign; *dynamic; election campaign*; environment; media; mobile media; *movement*; *network analysis*; news; online; online community; politic; *sentiment analysis*; social movement; television; united states
11	Information organization and diffusion	Information diffusion; organization; sentiment; social network	*Climate change*; information diffusion; organization; sentiment; social network
12	Network analysis and visualization	Big data; classification; data mining; detection; event detection; machine learning; microblog; retrieval; social sensing; topic detection; visualization	Big data; *classification; clustering; complex network; crowdsourcing*; data mining; detection; *emergency*; event detection; *feature; hashtag; location-based social network*; machine learning; microblog; *pattern*; retrieval; social sensing; topic detection; *unsupervised learning*; *visual analytic*; visualization
13	Emergent event management	Management	*Emergency management; extreme event*; management; Weibo
14	Adolescent online engagement	College student; Facebook; perception	*Adolescent*; college student; Facebook; *gender*; perception; response; *social networking site*

Note: Terms in italic type donate newly emerged keywords for period 2014–2017 compared with period 2009–2013

Table VIII. Potential research themes with keywords for the two periods

Social media has been widely adopted by emergency management organizations and agencies for disseminating emergency messages to the public (Ma and Yates, 2017). Thus, more and more researchers tend to make use of social media during emergency management and response (Pohl *et al.*, 2016). Adolescent online engagement focuses on adolescents' engagement in social media sites. Social media use among the child and

adolescent population is at an all-time high across the globe (Hur and Gupta, 2013). Researchers gradually pay attention to the use of adolescents' social media data for event detection such as psychological stress (Li, Xue, Zhao, Jia and Feng, 2017).

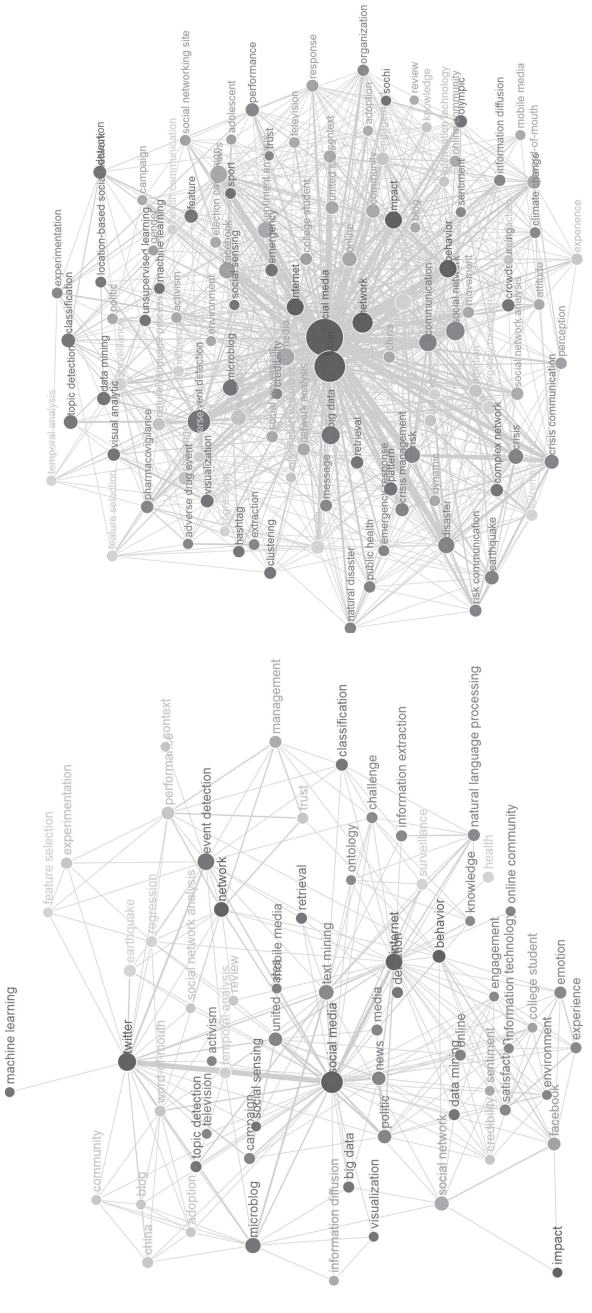
Visualization

We further conduct visualization of the results of keyword analysis and clustering analysis, shown in Figure 10. The node count and edge count for the first period are 65 and 257 with 2 sparse nodes communication and perception, while the node count and edge count for the second period are 110 and 1,000. In the visualized networks, a node represents a keyword while a link indicates the co-occurrence relation of a pair of keywords. The size of node is proportional to keyword frequency. The node color represents different clusters of the keywords while the link thickness indicates keyword co-occurrence relation strength. In the first period, when there is less research, only a few connections between keywords can be seen and the network is relatively dispersed. The top 5 keywords are social media, Twitter, event detection, *internet* and microblog. In the second period, there is a significant growth in keyword quantity as well as their relations. The top 5 keywords are social media, Twitter, event detection, network and social network. These largely correspond with the list of keywords in the results of the overall keyword co-occurrence analysis. The keywords for four themes, i.e. Twitter-related research, Text mining research, Blog community review and Online emotion detection, remain the same. This indicates that these themes are stable and are of continuing interest to scholars. Pharmacovigilance event detection and Sport related event are two newly emerged clusters in the second period. We here also provide interpretations for some of the newly emerged keywords. As for cluster 2, emerged keywords including crisis, crisis communication, crisis management, risk communication, etc., demonstrate that researchers gradually pay more attention to the use of social media data for crisis and risk communication and management. In cluster 7 for the second period, keywords adverse drug reaction, health communication and influenza indicate a new focus on adverse drug reaction and influenza ED in SM. For cluster 10, keywords attitude and sentiment analysis reveal an enthusiasm in identifying affective states and subjective attitude from social media data. As for cluster 12, keywords such as clustering, complex network, location-based social network, unsupervised learning and visual analytic represent some new techniques adopted to social media data analysis.

Discussion

This study provides an up-to-date bibliometric analysis on the publications of ED in SM research field in the period 2009–2017. The general publication distributions are identified to provide an overview of the current landscape of the research field. The increasing trends of both publication count and average citations indicate a growing interest and enthusiasm in the research field. Computer Science and Engineering are two major subjects that have contributed the most to the development of this field. *PLoS One* is the most productive journal while *IEEE Transactions on Multimedia* is the most influential one with the highest *H*-index. The USA and China are two most productive countries contributing for more than 50 percent of the total publications. Most productive countries are developed countries. ACP of internationally collaborated publications is much higher than that of non-internationally-collaborated publications for most of the highly productive countries/regions, indicating that international collaboration can improve publication quality in terms of citations. The top 4 affiliations ranked by publication quantity are all from China with Tsinghua University as the most productive one having the highest *H*-index. Xu, Zheng from China is the top 1 productive author with the highest *H*-index. The findings confirm that China is one active country in this research area.

The collaboration analysis indicates that authors tend to collaborate more with those from the same affiliation or country. This kind of collaboration can be affected by research subjects. Authors devoting to different subject demonstrate different levels of international



Sources: The dynamic visualization can be accessed via www.zhukun.org/haoty/resources.asp?id= JOIR_kword09-13 and www.zhukun.org/haoty/resources.asp?id= JOIR_kword14-17

Figure 10.
Keyword co-occurrence for the periods 2009–2013 (left) and 2014–2017 (right)

collaboration and affiliation collaboration. In terms of international collaboration, China has become the most important partner for the USA. The collaboration network reveals that affiliations tend to collaborate more with those from the same country. International collaboration among affiliations is limited and could be strengthened.

Through clustering analysis based on keyword co-occurrence, 14 major research themes are discovered, including Twitter-related research, Crisis communication research, Pharmacovigilance event detection, Sport related event, Experiment credibility, Text mining research, Health surveillance, Blog community review, Online emotion detection, etc. The network visualization of the findings through keyword analysis and clustering analysis for the two periods (2009–2013, 2014–2017) is also provided. Pharmacovigilance event detection and Sport related event are two newly emerged themes in the second period. Emerged keywords indicating new research interests include Crisis and risk communication and management, Adverse drug reaction and influenza event detection, Affective states and subjective attitude identification, Emergency management, etc.

We further compare our results with the findings from other existing bibliometric studies performed on social media-related topics, including social media-based innovation research (Appio *et al.*, 2016), social media research (Li, Wei, Xiong, Feng, Ye and Jiang, 2017), social media in hospitality and tourism research (Leung *et al.*, 2017), social media research with a focus on transport (Rashidi *et al.*, 2017) and social media in psychology research (Zyoud *et al.*, 2018). All the works have the similar reports on the growth of publications over time, demonstrating an increasing interest in social media-related research in academia.

For the analysis of productive journals, Leung *et al.* addressed that *Journal of Advertising* for business research and *Annals of Tourism Research* for hospitality research were the most prolific journals, while Appio *et al.* reported *Creativity and Innovation Management* as the top one. Li *et al.* and Zyoud *et al.* both reported that *Computers in Human Behavior* was the most prolific journal. However, in this study, the *Computers in Human Behavior* is ranked at 3rd while the top 1 is *PLoS One*. *PLoS One* was also the most productive one reported by Rashidi *et al.* For the analysis of productive countries, only Zyoud *et al.* and Li *et al.* reported the results, in which the USA and England were ranked at the top 2 in social media-related psychology research and social media research. In our study, the top prolific countries include the USA, China and England. The results indicate that the USA and England devote a lot to the development of social media-related research. For the analysis of major subjects, only Li *et al.* reported the result, in which Communication was ranked at the top 1. However, in our study, Communication is ranked at 3rd and the top 1 is Computer science. For the analysis of productive affiliations, only Zyoud *et al.* and Li *et al.* reported the results. The top 3 affiliations were all from the USA. Li *et al.* presented that University of Wisconsin was the most prolific affiliation with the most inter-institutional collaborative publications, while Zyoud *et al.* reported that University of Wisconsin-Madison was the top 1. Our study shows that the top 3 affiliations are all from China with Tsinghua University ranked at the top 1. The result reflects the change of academic research in the specific area.

For the analysis of keywords and topics, Leung *et al.* identified word-of-mouth as the major theoretical foundation of social media in business research. Li *et al.* demonstrated the research trends on human behavior and sustainability with keywords such as Facebook, Twitter, communication, social networking sites, China, climate change, big data and social support receiving increasing popularity. Rashidi *et al.* reported six most frequently used keywords including social media, Twitter, social networks, Facebook, data mining and location-based social networks. Our study reveals social media as the continuing foundation dominating in the center of the keyword networks. We also reports similar emerged keywords such as social networking site and climate change, and the top six frequently used keywords include social media, Twitter, event detection, network,

social network and internet. The results demonstrate that Twitter is a high concern in social media-related research and location and social network are essential focuses in transport-related social media research. Appio *et al.* reported five main research areas including Organizational learning, Open and distributed innovation, Value (co)creation, User/customer involvement in innovation processes and Knowledge sharing in communities. Zyoud *et al.* reported four most important research areas related to social media in psychology research including Personality psychology, Experimental psychology, Psychological risk factors and Developmental psychology. He mainly reported a concern for the developmental psychology of adolescent and college students as well as the risk factors such as alcohol for the young person or peer. Our study identifies 14 major research themes including Twitter-related research, Crisis communication research, Pharmacovigilance event detection, Sport related event, etc. We report a focus on teenagers' engagement in social networking site and find out the crisis, emergency, or natural disaster such as earthquake that may post threat to public health. The results indicate that learning, innovation, creation and knowledge sharing are important roles in social media-based innovation research, and social media in psychology research mainly focuses on psychology factor. However, ED in SM research demonstrates a wide concern for different specific areas.

In this study, we collect relevant publication data from WoS only. Though it is a widely applied repository for bibliometric analysis due to its high authority, some conference proceedings and journal articles may also be relevant to this field but have not been indexed in WoS. Thus, in our future work, we will consider including more academic research databases to collect more comprehensive data to avoid data bias. In the "Data collection" section, we conduct data verification to enhance the relevance of the data. Other data verification approaches for improving data retrieval quality can also be considered in the future.

In the analysis of research themes, author-defined keywords and ISI keywords-plus are used since they usually represent the main research focus of a publication. However, this might still lead to information loss without considering the content of the publication. Therefore, in future work, we will include natural language processing techniques such as word stemming, key phrase extraction and semantic relevance calculation for further improving the automated keyword matrix construction. The clustering analysis is performed based on keywords with a co-occurrence frequency $\geqslant 3$ to acquire a moderate category count. However, this might result in the ignorance of some sudden terms that are possible to represent research fronts although with low frequency. Therefore, we will try alternative methods such as Latent Dirichlet Allocation to consider every single term in our future work.

Conclusion

In this study, a bibliometric analysis on academic publications from WoS in the research field of ED in SM during the period 2009–2017 is conducted. We first identify the general publication distributions, including the trends of publications and citations, subject distribution, as well as productive journals, authors, affiliations, countries, etc. Annual collaboration degree distributions in the perspectives of country, affiliation and author are analyzed. The scientific collaboration relations are further visualized using a network analysis method. A clustering analysis based on a keyword co-occurrence analysis is applied to explore the major research themes. We further visualize the keyword analysis and clustering analysis results with a network analysis method to identify the emergence and evolution of keywords. The findings from the analysis, thus, comprehensively illustrate the research trend and development of ED in SM. They could potentially help relevant researchers understand the research status comprehensively, seek scientific collaborators and optimize research topic choices.

References

Ab Razak, M.F., Anuar, N.B., Salleh, R. and Firdaus, A. (2016), "The rise of 'Malware': bibliometric analysis of malware study", *Journal of Network and Computer Applications*, Vol. 75, pp. 58-76.

Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E. and Herrera, F. (2009), "*H*-Index: a review focused in its variants, computation and standardization for different scientific fields", *Journal of Informetrics*, Vol. 3 No. 4, pp. 273-289.

Amato, F., Moscato, V., Picariello, A., Piccialli, F. and Sperli, G. (2018), "Centrality in heterogeneous social networks for Lurkers detection: an approach based on hypergraphs", *Concurrency and Computation-Practice & Experience*, Vol. 30 No. 3, pp. 1-12.

Appio, F.P., Martini, A., Massa, S. and Testa, S. (2016), "Unveiling the intellectual origins of social media-based innovation: insights from a bibliometric approach", *Scientometrics*, Vol. 108 No. 1, pp. 355-388.

Atefeh, F. and Khreich, W. (2015), "A survey of techniques for event detection in Twitter", *Computational Intelligence*, Vol. 31 No. 1, pp. 132-164.

Atkinson, C.L. (2014), "Crisis communication in dark times: the 2011 mouse river flood in Minot, North Dakota", *International Journal of Communication*, Vol. 8 No. 2, pp. 1394-1414.

Bouyssou, D. and Marchant, T. (2011), "Ranking scientists and departments in a consistent manner", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 9, pp. 1761-1769.

Bravo-Marquez, F., Mendoza, M. and Poblete, B. (2014), "Meta-level sentiment models for big social data analysis", *Knowledge-Based Systems*, Vol. 69, pp. 86-99.

Chen, X.L., Ding, R.Y., Xu, K., Wang, S., Hao, T.Y. and Zhou, Y. (2018), "A bibliometric review of natural language processing empowered mobile computing", *Wireless Communications and Mobile Computing*, Vol. 2018 No. SI, pp. 1-21.

Chen, X.L., Xie, H.R., Wang, F.L., Liu, Z.Q., Xu, J. and Hao, T.Y. (2018), "A bibliometric analysis of natural language processing in medical research", *BMC Medical Informatics and Decision Making*, Vol. 18 No. S1, pp. 1-14.

Cheng, T. and Wicks, T. (2014), "Event detection using Twitter: a spatio-temporal approach", *Plos One*, Vol. 9 No. 6, pp. 1-10.

Chiu, W.T. and Ho, Y.S. (2007), "Bibliometric analysis of Tsunami research", *Scientometrics*, Vol. 73 No. 1, pp. 3-17.

Cobo, M.J., Martinez, M.A., Gutierrez-Salcedo, M., Fujita, H. and Herrera-Viedma, E. (2015), "25 years at knowledge-based systems: a bibliometric analysis", *Knowledge-Based Systems*, Vol. 80, pp. 3-13.

Dong, X.W., Mavroeidis, D., Calabrese, F. and Frossard, P. (2015), "Multiscale event detection in social media", *Data Mining and Knowledge Discovery*, Vol. 29 No. 5, pp. 1374-1405.

Dou, W., Wang, X., Ribarsky, W. and Zhou, M. (2012), "Event detection in social media data", IEEE VisWeek Workshop on Interactive Visual Text Analytics-Task Driven Analytics of Social Media Content, pp. 971-980.

Gao, C. and Liu, J.M. (2017), "Network-based modeling for characterizing human collective behaviors during extreme events", *IEEE Transactions on Systems Man Cybernetics-Systems*, Vol. 47 No. 1, pp. 171-183.

Harpaz, R., Callahan, A., Tamang, S., Low, Y., Odgers, D., Finlayson, S., Jung, K., LePendu, P. and Shah, N.H. (2014), "Text mining for adverse drug events: the promise, challenges, and state of the art", *Drug Safety*, Vol. 37 No. 10, pp. 777-790.

Hirsch, J.E. (2005), "An index to quantify an individual's scientific research output", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 102 No. 46, pp. 16569-16572.

Hua, T., Chen, F., Zhao, L., Lu, C.T. and Ramakrishnan, N. (2016), "Automatic targeted-domain spatiotemporal event detection in Twitter", *Geoinformatica*, Vol. 20 No. 4, pp. 765-795.

Hur, J.L. and Gupta, M. (2013), "Growing up in the web of social networking: adolescent development and social media", *Adolescent Psychiatry*, Vol. 3 No. 3, pp. 233-244.

Jiang, D.D., Luo, X.F., Xuan, J.Y. and Xu, Z. (2017), "Sentiment computing for the news event based on the social media big data", *IEEE Access*, Vol. 5, pp. 2373-2382.

Juliani, F. and de Oliveira, O.J. (2016), "State of research on public service management: identifying scientific gaps from a bibliometric study", *International Journal of Information Management*, Vol. 36 No. 6, pp. 1033-1041.

Kaleel, S.B. and Abhari, A. (2015), "Cluster-discovery of Twitter messages for event detection and trending", *Journal of Computational Science*, Vol. 6, pp. 47-57.

Lee, P.C. and Su, H.N. (2010), "Investigating the structure of regional innovation system research through keyword co-occurrence and social network analysis", *Innovation-Management Policy & Practice*, Vol. 12 No. 1, pp. 26-40.

Leung, X.Y., Sun, J. and Bai, B. (2017), "Bibliometrics of social media research: a co-citation and co-word analysis", *International Journal of Hospitality Management*, Vol. 66, pp. 35-45.

Li, C., Sun, A. and Datta, A. (2012), "Twevent: segment-based event detection from tweets", *The 21st ACM International Conference on Information and Knowledge Management, ACM*, pp. 155-164.

Li, Q., Xue, Y.Y., Zhao, L., Jia, J. and Feng, L. (2017), "Analyzing and identifying teens' stressful periods and stressor events from a microblog", *IEEE Journal of Biomedical and Health Informatics*, Vol. 21 No. 5, pp. 1434-1448.

Li, Q., Wei, W.B., Xiong, N.A., Feng, D.C., Ye, X.Y. and Jiang, Y.S. (2017), "Social media research, human behavior, and sustainable society", *Sustainability*, Vol. 9 No. 3, pp. 1-11.

Liu, Z.G., Yin, Y.M., Liu, W.D. and Dunford, M. (2015), "Visualizing the intellectual structure and evolution of innovation systems research: a bibliometric analysis", *Scientometrics*, Vol. 103 No. 1, pp. 135-158.

Lorentzen, D.G. (2016), "Twitter conversation dynamics of political controversies: the case of Sweden's December agreement", *Information Research-an International Electronic Journal*, Vol. 21 No. 2, p. 1.

Luo, Q.J. and Zhai, X.T. (2017), "'I will never go to Hong Kong Again!' How the secondary crisis communication of 'Occupy Central' on Weibo Shifted to a Tourism Boycott", *Tourism Management*, Vol. 62, pp. 159-172.

Ma, X. and Yates, J. (2017), "Multi-network multi-message social media message dissemination problem for emergency communication", *Computers & Industrial Engineering*, Vol. 113, pp. 256-268.

Mazloumian, A. (2012), "Predicting scholars' scientific impact", *Plos One*, Vol. 7 No. 11, pp. 1-5.

Merigo, J.M., Gil-Lafuente, A.M. and Yager, R.R. (2015), "An overview of fuzzy research with bibliometric indicators", *Applied Soft Computing*, Vol. 27, pp. 420-433.

Nawaz, M.S., Bilal, M., Lali, M.I., Ul Mustafa, R., Aslam, W. and Jajja, S. (2017), "Effectiveness of social media data in healthcare communication", *Journal of Medical Imaging and Health Informatics*, Vol. 7 No. 6, pp. 1365-1371.

Nurwidyantoro, A. and Winarko, E. (2013), "Event detection in social media: a survey", *2013 International Conference on ICT for Smart Society*, pp. 307-311.

Petrović, S., Osborne, M. and Lavrenko, V. (2010), "Streaming first story detection with application to Twitter", *Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, CA*, pp. 181-189.

Petz, G., Karpowicz, M., Furschuss, H., Auinger, A., Stritesky, V. and Holzinger, A. (2015), "Computational approaches for mining user's opinions on the web 2.0", *Information Processing & Management*, Vol. 51 No. 4, pp. 510-519.

Petz, G., Karpowicz, M., Furschuss, H., Auinger, A., Winkler, S.M., Schaller, S. and Holzinger, A. (2012), "On text preprocessing for opinion mining outside of laboratory environments", *International Conference on Active Media Technology in Berlin, Heidelberg, Springer*, pp. 618-629.

Pohl, D., Bouchachia, A. and Hellwagner, H. (2016), "Online indexing and clustering of social media data for emergency management", *Neurocomputing*, Vol. 172, pp. 168-179.

Pu, Q.H., Lyu, Q.J. and Su, H.Y. (2016), "Bibliometric analysis of scientific publications in transplantation journals from Mainland China, Japan, South Korea and Taiwan between 2006 and 2015", *BMJ Open*, Vol. 6 No. 8, pp. 1-7.

Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S. and Waller, T.S. (2017), "Exploring the capacity of social media data for modelling travel behaviour: opportunities and challenges", *Transportation Research Part C: Emerging Technologies*, Vol. 75, pp. 197-211.

Rey-Marti, A., Ribeiro-Soriano, D. and Palacios-Marques, D. (2016), "A bibliometric analysis of social entrepreneurship", *Journal of Business Research*, Vol. 69 No. 5, pp. 1651-1655.

Serrat, O. (2017), "Social network analysis", in Serrat, O. (Ed.), *Knowledge Solutions: Tools, Methods, and Approaches to Drive Organizational Performance*, Springer, Singapore, pp. 39-43.

Srijith, P.K., Hepple, M., Bontcheva, K. and Preotiuc-Pietro, D. (2017), "Sub-story detection in twitter with hierarchical Dirichlet processes", *Information Processing & Management*, Vol. 53 No. 4, pp. 989-1003.

Stephen, A.T. and Galak, J. (2012), "The effects of traditional and social earned media on sales: a study of a micro lending marketplace", *Journal of Marketing Research*, Vol. 49 No. 5, pp. 624-639.

Taraghi, B., Grossegger, M., Ebner, M. and Holzinger, A. (2013), "Web analytics of user path tracing and a novel algorithm for generating recommendations in open journal systems", *Online Information Review*, Vol. 37 No. 5, pp. 672-691.

Wang, D., Al-Rubaie, A., Clarke, S.S. and Davies, J. (2017), "Real-time traffic event detection from social media", *ACM Transactions on Internet Technology*, Vol. 18 No. 1, pp. 1-23.

Wei, Y.M., Mi, Z.F. and Zhang, H. (2013), "Progress of integrated assessment models for climate policy", *Systems Engineering-Theory & Practice*, Vol. 33 No. 8, pp. 1905-1915.

Wei, Y.M., Yuan, X.C., Wu, G. and Yang, L.X. (2014), "Climate change risk assessment: a bibliometric analysis based on web of science", *Bulletin of National Natural Science Foundation of China*, Vol. 28 No. 5, pp. 347-356.

Weiler, A., Grossniklaus, M. and Scholl, M.H. (2016), "Situation monitoring of urban areas using social media data streams", *Information Systems*, Vol. 57, pp. 129-141.

Xu, X., Mishra, G.D. and Jones, M. (2017), "Mapping the global research landscape and knowledge gaps on multimorbidity: a bibliometric study", *Journal of Global Health*, Vol. 7 No. 1, pp. 1-11.

Xuan, J., Luo, X., Lu, J. and Zhang, G. (2017), "Explicitly and implicitly exploiting the hierarchical structure for mining website interests on news events", *Information Sciences*, Vol. 420, pp. 263-277.

Yang, H.D. and Yang, C.C. (2015), "Using health-consumer-contributed data to detect adverse drug reactions by association mining with temporal analysis", *ACM Transactions on Intelligent Systems and Technology*, Vol. 6 No. 4, pp. 1-27.

Yeleswarapu, S.J., Rao, A., Joseph, T. and Srinivasan, R. (2014), "A pipeline to extract drug-adverse event pairs from multiple data sources", *BMC Medical Informatics and Decision Making*, Vol. 14 No. 1, pp. 1-16.

Yu, D., Xu, Z. and Wang, W. (2018), "Bibliometric analysis of fuzzy theory research in China: a 30-year perspective", *Knowledge-Based Systems*, Vol. 141, pp. 188-199.

Zhong, S., Geng, Y., Liu, W., Gao, C. and Chen, W. (2016), "A bibliometric review on natural resource accounting during 1995-2014", *Journal of Cleaner Production*, Vol. 139, pp. 122-132.

Zyoud, S.H., Sweileh, W.M., Awang, R. and Al-Jabi, S.W. (2018), "Global trends in research related to social media in psychology: mapping and bibliometric analysis", *International Journal of Mental Health Systems*, Vol. 12 No. 1, pp. 1-8.

Further reading

Garfield, E. (2006), "The history and meaning of the journal impact factor", *JAMA-Journal of the American Medical Association*, Vol. 295 No. 1, pp. 90-93.

Corresponding author
Tianyong Hao can be contacted at: haoty@126.com

What the fake? Assessing the extent of networked political spamming and bots in the propagation of #fakenews on Twitter

Received 2 March 2018
Revised 7 July 2018
18 September 2018
Accepted 19 September 2018

Ahmed Al-Rawi
Simon Fraser University, Burnaby, Canada, and
Jacob Groshek and Li Zhang
Boston University College of Communication, Boston, Massachusetts, USA

Abstract

Purpose – The purpose of this paper is to examine one of the largest data sets on the hashtag use of #fakenews that comprises over 14m tweets sent by more than 2.4m users.
Design/methodology/approach – Tweets referencing the hashtag (#fakenews) were collected for a period of over one year from January 3 to May 7 of 2018. Bot detection tools were employed, and the most retweeted posts, most mentions and most hashtags as well as the top 50 most active users in terms of the frequency of their tweets were analyzed.
Findings – The majority of the top 50 Twitter users are more likely to be automated bots, while certain users' posts like that are sent by President Donald Trump dominate the most retweeted posts that always associate mainstream media with fake news. The most used words and hashtags show that major news organizations are frequently referenced with a focus on CNN that is often mentioned in negative ways.
Research limitations/implications – The research study is limited to the examination of Twitter data, while ethnographic methods like interviews or surveys are further needed to complement these findings. Though the data reported here do not prove direct effects, the implications of the research provide a vital framework for assessing and diagnosing the networked spammers and main actors that have been pivotal in shaping discourses around fake news on social media. These discourses, which are sometimes assisted by bots, can create a potential influence on audiences and their trust in mainstream media and understanding of what fake news is.
Originality/value – This paper offers results on one of the first empirical research studies on the propagation of fake news discourse on social media by shedding light on the most active Twitter users who discuss and mention the term "#fakenews" in connection to other news organizations, parties and related figures.

Keywords Twitter, Fake news, Bots, Networked political spamming

Paper type Research paper

Introduction

This study sheds light on the most active Twitter users who discuss and mention the term "#fakenews" in connection to other news organizations, parties and related figures. It also investigates whether these users are more likely to be humans or bots in order to better understand the nature of the dissemination of discourses surrounding fake news discussion on social media. In this regard, there is also another category called cyborg that combines both artificial and human activity. For example, Daniel John Sobieski, a Conservative Activist on Twitter with the username @gerfingerpoken, uses algorithms to post over 1,000 messages a day in order to further his agenda and reach a wider online public. This is just one of the actions that cyborgs can provide, and in this case Sobieski uses "schedulers" which "work through stacks of his own prewritten posts in repetitive loops" (Timberg, 2017). Further, "political bots tend to be developed and deployed in sensitive political

This paper forms part of a special section "Social media mining for journalism".

Online Information Review
Vol. 43 No. 1, 2019
pp. 53-71

1468-4527
DOI 10.1108/OIR-02-2018-0065

moments when public opinion is polarized" (Kollanyi *et al.*, 2016, p. 1). For example, one study on Twitter found that "almost 50% of traffic is generated and propagated by a rapidly growing bot population" (Gilani *et al.*, 2017).

In the contemporary media environment, fake news is becoming more important than perhaps ever before as "political actors and governments worldwide have begun using bots to manipulate public opinion, choke off debate, and muddy political issues" (Forelle *et al.*, 2015, p. 1). Indeed, fake news has become a highly partisan issue in the USA, so associating certain political figures or news organizations with making or spreading it can lead to undermining their credibility. This study attempts to examine the way some active Twitter users connect certain figures, parties and sides with fake news, which can be regarded as a part of their political spamming activities that are meant to discredit their ideological opponents. There is no doubt that there is an increasing interest by the general public in the issue of fake news especially due to its importance in influencing campaigns, shaping the perception of reality and potentially altering citizens' political decision making. In general, there seems to be a systematic and well-calculated attack on mainstream media by many political sides in the way it is associated with fake news (Cadwalladr, 2017).

The main issue here is that most social media sites like Twitter and Facebook allow bots to be used, which boost and enhance spamming or posting messages by repeatedly sending them to as many other users as possible (Chu *et al.*, 2010). For example, Donald Trump's first presidential address was initially identified as the most tweeted event in history, but it has been observed that this online attention was partly due to the use of pro-Trump bots. To wit, "Even before they started trending […], the official hashtags – #JointAddress and #JointSession – accumulated decidedly inorganic traffic, including from some accounts that had never tweeted about any other topic" (Musgrave, 2017). Some of these accounts are not totally automated as there seems to be cyborgs or human spammers and bot activity as explained above, for such "accounts are often bots that see occasional human curation, or they are actively maintained by people who employ scheduling algorithms and other applications for automating social media communication" (Kollanyi *et al.*, 2016, p. 2). According to Pew Research Center, it has been estimated that two-thirds of "tweeted links to popular websites are posted by" bots that "share roughly 41% of links to political sites shared primarily by liberals and 44% of links to political sites shared primarily by conservatives" (Wojcik *et al.*, 2018).

Theoretical framework

Since this study deals with online information, it is relevant to begin with the theoretical concept of political spamming, which we define as an overflow of politically oriented online messages that are widely disseminated to serve the interest of a certain political party or figure. In the context of this study, spamming is done with the way news organizations, political figures and entities are repeatedly associated with fake news on Twitter. Further, we introduce here the concept of networked political spamming activity which is manifested in the way many, active Twitter users collaboratively disseminate posts by retweeting political or ideological messages that often include hyperlinks in order to serve a certain agenda or political purpose. The majority of previous studies on political spamming did not offer a clear conceptual definition of this online activity, while the networked and collaborative aspect has been largely overlooked. This is a networked activity because there is a collective collaboration in disseminating spam, and those involved might not always be aware of their spamming activity. Though spam is not always defined as a form of false information, it is somehow similar to the spread of misinformation which refers to the "inadvertent sharing" of wrong information when users are not aware of the nature of messages they disseminate (Born and Edgington, 2017; Jackson, 2017). In other words, networked political spamming includes the intentional and unintentional spread of spam messages by social media users whose general aim is to serve a particular political side and attack or silence the opponent(s).

In general, spamming is not a new phenomenon in politics. For example, during the time fax machines were still popular in the 1990s, a US company called Bonner and Associates was "able to send out 10,000 faxes overnight to a congressperson's office. When the firm is hired by a client, it isolates the 'swing votes' in Congress, does a scan of the corresponding districts, and identifies citizens whose profiles suggest that they are sympathetic to the cause" (Newman, 1999, p. 6). Other types of spam include commercial ones, pre-recorded telephone messages and snail mail.

As for online spamming, it has been mostly done through e-mails to achieve unconventional political mobilization purposes, and it is considered a much cheaper option than political advertising on TV or radio (Sweet, 2003; Krueger, 2010). Online political spamming has become part of the new political reality. For example, during the 2002 US midterm elections, many politicians from different political affiliations sent voters many unsolicited e-mails, widely regarded as unregulated political speech (Sweet, 2003). In relation to political campaign activities in the year 2004, "it was estimated that over 1.25 billion political spam messages were sent during the campaign as many candidates used e-mail to supplement direct mail campaigns" (Quinn and Kivijarv, 2005, p. 136). However, the actual impact of such a strategy is uncertain as it is mostly expected to influence swing voters (Frankel and Hillygus, 2014, p. 184) and is widely regarded as a "bad politics" strategy (Krueger, 2006, p. 763). It is believed that "large scale political spamming" usually done through e-mails is unethical and can have a negative impact on democracy and political deliberation, so there should be some kind of regulation to control its impact on citizens (Grossman, 2004; Rooksby, 2007; Treré, 2016), while other scholars think that political spamming should be protected as part of the First Amendment (Sweet, 2003).

In relation to Twitter, spamming occurs in the way certain political campaigns are implemented and messages are repeatedly retweeted often with the use of cyborgs and bots (Gao *et al.*, 2010; Sridharan *et al.*, 2012). It also works by including hyperlinks in the tweets "that a user would likely not visit otherwise" (Just *et al.*, 2012, p. 16). Though a few previous studies showed that political spam was not prevalent on Twitter during the 2008 US Congressional Elections or in the discussion of certain controversial political topics (Metaxas and Mustafaraj, 2009; Himelboim *et al.*, 2013), this research argues, based on the empirical findings, that political spamming is very prevalent in the context of discourse on fake news.

This position is in line with many other studies conducted on the Twitter spam use during the 2010 municipal elections in Ottawa, Canada (Raynauld and Greenberg, 2014) and the Massachusetts (MA) senate race between Martha Coakley and Scott Brown in 2010 (Mustafaraj and Metaxas, 2010). In relation to the latter elections, many spammers targeted "individual journalists and liberal media outlets" in order to discredit them (Just *et al.*, 2012), and the examination of the top 200 most active accounts revealed that a small number of users attempted to game search engines as they "were responsible for many of the replies, in an attempt to flood the network with spam" (Mustafaraj and Metaxas, 2010, p. 2). Several other studies showed similar results on the impact of spamming on political deliberation and debates. For example, Verkamp and Gupta (2013) studied the popular hashtags around five political protests and events from 2011 and 2012 from different parts of the world and found that the hashtags were "inundated with spam tweets intended to overwhelm the original content" in an attempt to silence dissent.

As mentioned above, there is a gap in literature with regard to empirically studying fake news, specifically fake news discourses whose audiences can be vastly increased by spammers, whether be bots or humans. On this front, there are some studies that have examined bots during Brexit (Howard and Kollanyi, 2016; Gallacher *et al.*, 2017) and the 2016 election (Kollanyi *et al.*, 2016). These researchers examined election hashtags and found that "Twitter traffic on pro-Trump hashtags was roughly double that of the pro-Clinton hashtags, and about one third of the pro-Trump twitter traffic was driven by bots and

highly automated accounts, compared to one-fifth of the pro-Clinton twitter traffic." Similarly, Bessi and Ferrara (2016) found that about 19 percent of all 2016 US election tweets were sent by political bots, amounting to about one-fifth of the total communication on Twitter related to this topic.

Given the opaque and still-debated scope of fake news and the most-influential users referencing fake news, this study attempts to provide an understanding of fake news discussion on social media. While such an endeavor cannot prove effects of exposure to fake news, it very well can provide vital insights about fake news as a cultural phenomenon as it is debated on social media. As such, we therefore pose the following research questions:

RQ1. In relation to networked political spamming, what are the most associated users and hashtags that are linked to #fakenews mentions on Twitter as well as the most retweeted posts?

RQ2. Which Twitter accounts are the most active in spamming and disseminating #fakenews tweets, and what is the likelihood that they are bots?

Methods

In order to identify an appropriate time frame in which to study fake news, we rely on data from Google and Wikipedia search traffic. Here, according to a Google Trend search, the term "fake news" became popular online in January 2017, which corresponds with the highlighting of this term on various topics by the current US President, Donald Trump, and many other politicians, journalists, and the general public following the US elections (see Figure 1). The highest peak in Google searches for this term was, in fact, in mid-January 2018 when Donald Trump announced his fake news awards contest (Siddiqui, 2018). This denotes the way famous figures like the US President can popularize certain terms. On Wikipedia, the highest number of searches for the "fake news" entry occurred between October and November 2017 (see Figure 2). Since these are important periods for researching fake news, we have chosen to study this topic around these dates.

Our data on fake news tweets were collected from the Boston University Twitter Collection and Analysis Toolkit, where data collection remains ongoing (Borra and Rieder, 2014; Groshek, 2014). As indicated above, Google and Wikipedia searches indicate that there has been an increasing public interest in this topic starting from January and February 2017, so we collected tweets on the hashtag (#fakenews) for a period of over one year from January 3 to May 7 of 2018. In total, there were 14,300,463 tweets retrieved that were posted by 2,493,949 unique users, and the highest peak of tweets was found in January 11 with 151,735 units collected that day. On the whole, 49.6 percent of tweets have links to other sites (see Figure 3).

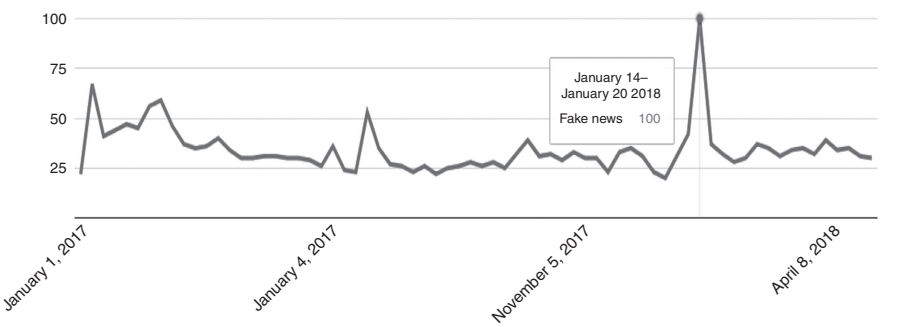


Figure 1.
Google searches for "fake news" from January to May 2018

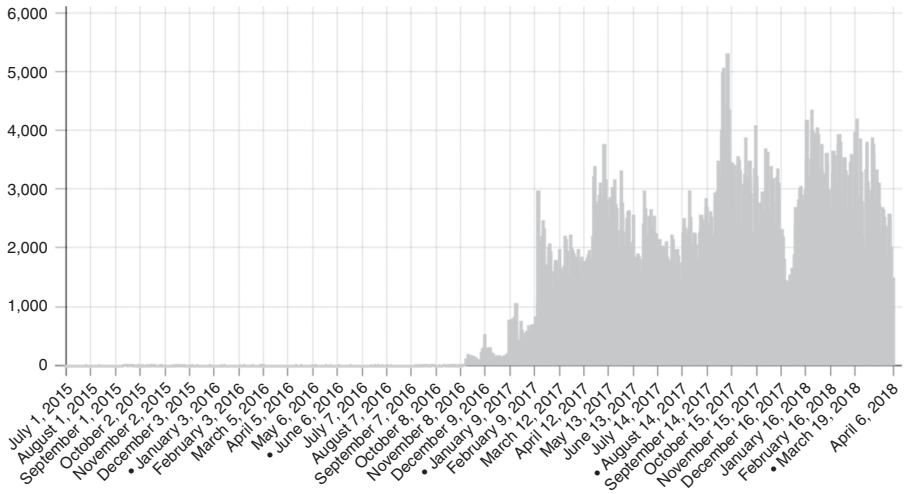


Figure 2.
Wikipedia searches for "fake news" from July 2015 to May 2018

In the second stage of the study, we examined the most mentioned terms associated with the hashtag #fakenews on Twitter as well as the top 50 most active users in terms of the frequency of their tweets. Since there is a lot of noise and irrelevant content on social media, the choice was to select the top 50 users following previous research that examined large data sets (Wilkinson and Thelwall, 2012; Al-Rawi, 2017a, b). We used Gephi (https://gephi.org/), an open source visualization software (Bastian *et al.*, 2009), in order to present a graph that models the influence and communities around the most mentioned users and their connections with other users mentioning each other in the network constructed around this topic. To take on additional analytic step, we used an online tool called botometer (https://botometer.iuni.iu.edu) in order to understand the bots' scores of the top Twitter accounts (Davis *et al.*, 2016; Bessi and Ferrara, 2016; Shao *et al.*, 2017; Ferrara, 2017, 2018). Previous research showed the effectiveness of this award-winning tool, and it can be regarded as a useful starting point for an exploratory study such as this.

The above methods are relevant in understanding the Twitter users that most actively spread information on fake news, their affiliations, the nature of such accounts in terms of being a bot or human. As far as the researchers' knowledge, this is the first empirical study that examines fake news using the above methodological procedures, which can altogether assist in filling an important gap in literature and advance future understanding of a growing sociopolitical concern.

Findings

To answer the first research question on the most associated @usernames that are linked to #fakenews mentions on Twitter, it can be observed that @realdonaldtrump with 1,330,141 such mentions ranks first followed by @CNN $n = 1{,}164{,}871$ mentions followed by @potus (President of the USA) at 472,656, @nytimes with 212,092 and @foxnews with 209,476 mentions on Twitter. There is clear tendency toward Twitter handles that represent media organizations or politicians that can be further illustrated in the following ranking of mentions: (6) @donaldjtrumpjr $n = 201{,}600$, (7) @msnbc $n = 145{,}621$, (8) @washingtonpost $n = 144{,}717$, (16) @abc $n = 103{,}675$, (17) @nbcnews $n = 90{,}838$, (28) @thehill $n = 52{,}505$, (32) @cnnpolitics $n = 48{,}118$, (35) @cbsnews $n = 46{,}885$, (36) @nbc $n = 46{,}705$, (37) @hillaryclinton $n = 45{,}186$, (38) @cbs $n = 45{,}135$ and (41) @ap $n = 39{,}977$.

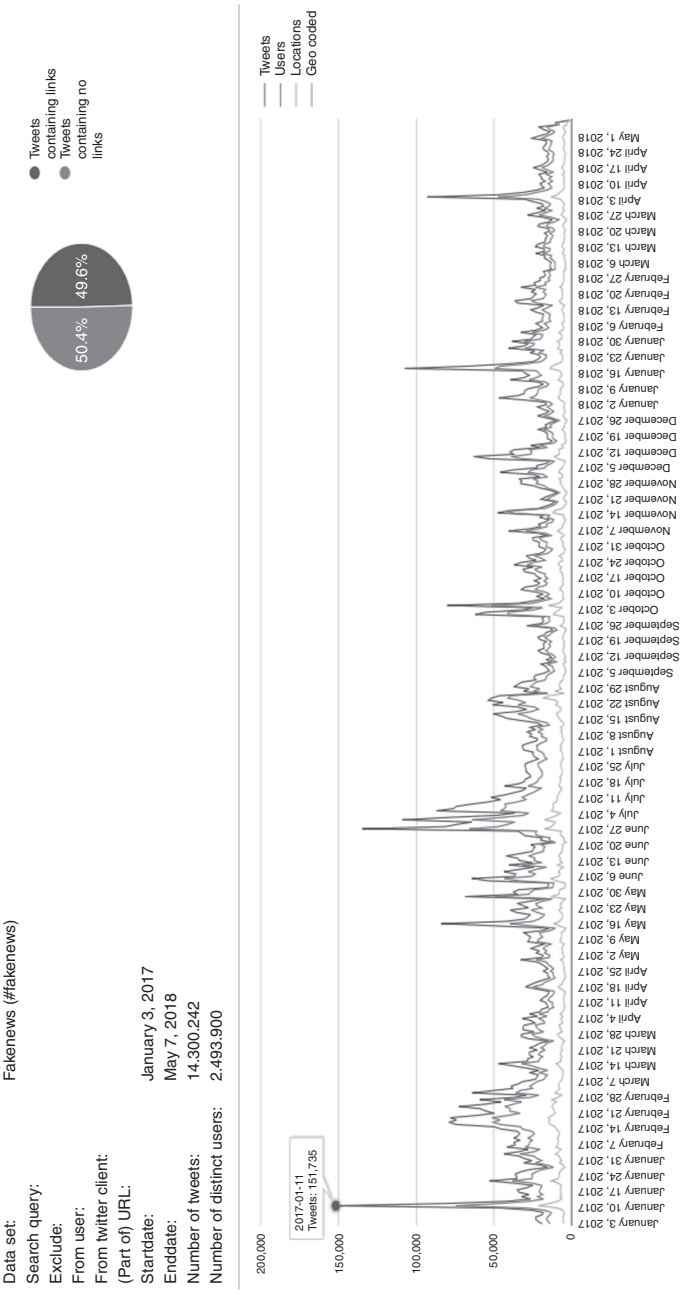


Figure 3.
Distribution of tweets mentioning #fakenews from January 3, 2017 to May 7, 2018

Further, there are many other mentions of journalists working for media outlets that were referenced very frequently, specifically including Jake Tapper (CNN; $n = 76{,}210$), Jim Acosta (CNN; $n = 47{,}023$), Chris Cuomo (CNN; $n = 111{,}767$) and Brian Stelter (CNN; $n = 38{,}492$). However, there are many other references to users (either human or bot) that are linked to or supportive of Donald Trump, such as James Woods ($n = 142{,}020$), Bill Mitchell ($n = 134{,}643$), the host of YourVoice at www.yourvoiceamerica.tv, Kevin W. ($n = 1{,}27{,}502$) James Edward O'Keefe III ($n = 122{,}752$) and Linda Suhler ($n = 107{,}359$) who is regarded as one of "Trump's female internet superfans" (Roller, 2016) but is believed to be an account with almost exclusively bot-like behavior (Bohannon, 2017) (see Table I).

Beyond the simple frequency of user mentions, we also used network analysis to construct a social graph by mentions to identify especially influential users and communities of users with the network of discussion on this topic. Here, weighted degree metrics were used to size user nodes and thereby determine their influence in spreading messages through the network by their activity in mentioning and being mentioned by other users of influence. The modularity algorithm placed users into communities within this network and is identified by color in the graph. The most active 1,500 nodes (connected with 56,505 edges) were spatialized using the Open Ord algorithm in Gephi, which is suitable to better distinguish clusters of users.

As shown in Figure 4, this graph is also available online in a dynamic interactive user interface at https://bit.ly/2zcIraL. When sorted by user influence, many of the same accounts appeared in this graph, specifically with the top 20 being @cnn, @potus, @realdonaldtrump, @trey_vondinkis, @deplorable80210, @siddonsdan, @americanvoterus, @kwilli1046, @jrcheneyjohn, @rodstryker, @lawriter33, @rosenchild, @drmartyfox, @jimiznhb, @nytimes, @lvnancy, @georgiadirtroad, @poetreeotic, @petefrt and @msnbc.

Though there is no simple obvious pattern, the majority of tweets reference mainstream media as there seems to be a systematic and ongoing identifications with mainstream news

Rank	Mention	Frequency	Rank	Mention	Frequency
1.	realdonaldtrump	1,330,141	26.	christichat	56,572
2.	cnn	1,164,871	27.	drmartyfox	52,675
3.	potus	472,656	28.	thehill	52,505
4.	nytimes	212,092	29.	ingrahamangle	52,081
5.	foxnews	209,476	30.	teapainusa	49,895
6.	donaldjtrumpjr	201,600	31.	bfraser747	48,873
7.	msnbc	145,621	32.	cnnpolitics	48,118
8.	washingtonpost	144,717	33.	presssec	47,785
9.	realjameswoods	142,020	34.	johncardillo	47,095
10.	mitchellvii	134,643	35.	cbsnews	46,885
11.	kwilli1046	127,502	36.	nbc	46,705
12.	jamesokeefeiii	122,752	37.	hillaryclinton	45,186
13.	jaketapper	111,767	38.	cbs	45,135
14.	lindasuhler	107,359	39.	markdice	43,253
15.	project_veritas	104,271	40.	gmoneyrainmaker	42,946
16.	abc	103,675	41.	ap	39,977
17.	nbcnews	90,838	42.	wikileaks	39,238
18.	acosta	89,311	43.	realalexjones	38,558
19.	seanhannity	88,071	44.	brianstelter	38,492
20.	sebgorka	75,167	45.	donlemon	37,028
21.	georgiadirtroad	71,556	46.	_makada_	35,885
22.	jrcheneyjohn	66,822	47.	sandratxas	35,734
23.	lvnancy	66,003	48.	lkirchner	35,615
24.	americanvoterus	63,048	49.	thejefflarson	35,615
25.	bocavista2016	56,920	50.	repstevensmith	34,864

Table I. The top 50 most mentions in connection with #fakenews

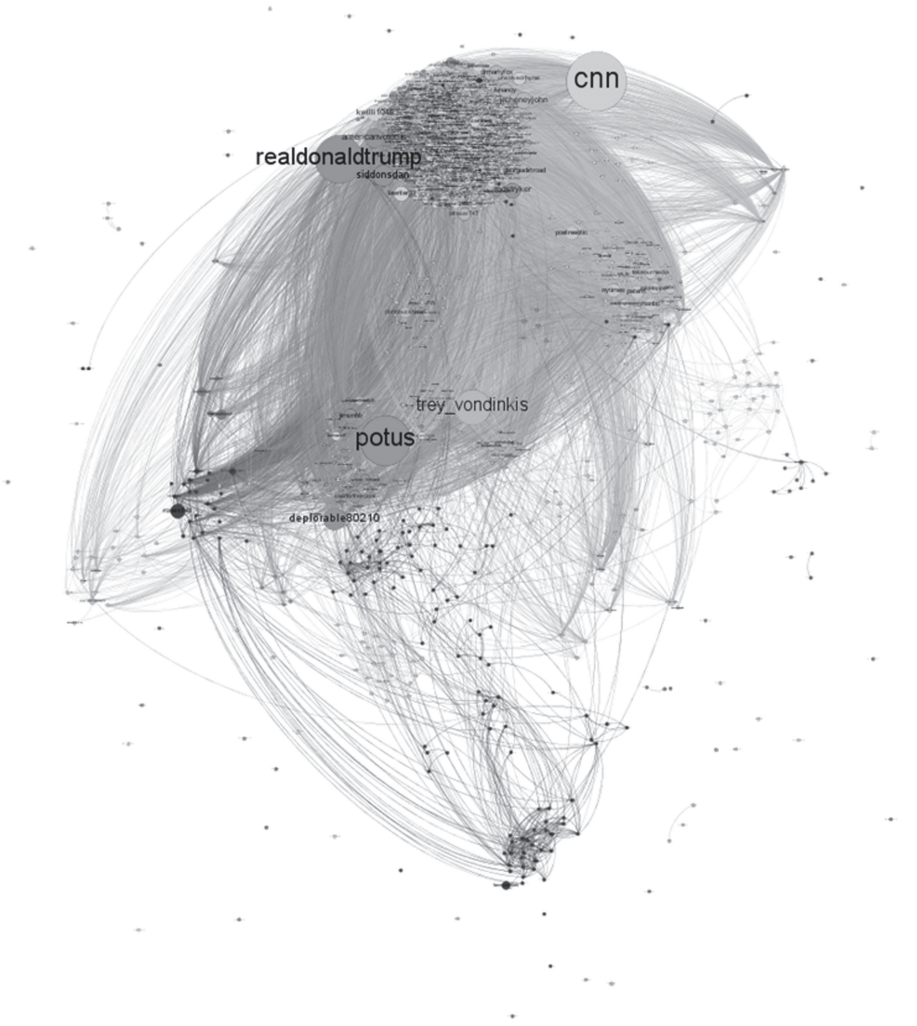


Figure 4.
A social network graph by mentions identifying the most active 1,500 nodes connected with 56,505 edges[a]

Note: [a]For a higher resolution and more detailed graph, see the following link: https://bit.ly/2zcIraL

organizations, especially CNN, which is by far the most mentioned outlet. There is evidence suggesting that the current US President and many members of conservative, Republican and (far) right groups have been involved in attacking CNN, identifying it as fake news as explained below. This partly explains the high frequency of mentions to this news channel. However, many other users reference CNN in connection to fake news in order to defend it rather than attack the news outlet. By examining references to the names of other news organizations, we find that the overwhelming majority are considered liberal such as CBS, MSNBC, NBC, *NYT* and *Washington Post*, while only one is typically regarded as conservative, namely Fox News.

As can be seen above, the most mentioned news outlet that is often associated with #fakenews references is CNN. In order to dig deeper into the data and understand how CNN is connected, we further examined the most used hashtags in the data set. Aside from the common ones, some of which are already covered above in the reporting on the most mentions such as #CNN (rank 4, $n = 439{,}095$), we find that there are certain terms in the top 50 most used hashtags that are clearly negative such as #FakeNewsCNN (rank 10, $n = 106{,}168$), #CNNBlackmail (rank 13, $n = 85{,}265$), #FraudNewsCNN (rank 31, $n = 51{,}470$) and #CnnIsIsis (rank 49, $n = 33{,}307$). In other words, CNN is mostly associated with negative terms that are connected to fake news discourses probably to undermine its credibility and status as a well-known mainstream media outlet (Tables I and II).

In order to further understand the most prevalent messages found in the data set, we investigated the top 25 most retweeted posts, which were retweeted 506,944 times in total. This examination is important and relevant because it provides an indication into the kind of messages. Twitter users are mostly engaged with and interested in retweeting. Once more, we find that CNN is the most referenced news outlet ($n = 5$) followed by ProPublica ($n = 2$) and NBC ($n = 2$) that are all framed in a negative manner, while Donald Trump @realDonaldTrump and his son @DonaldJTrumpJr have dominated the online chatter with 16 tweets that were retweeted 315,140 times, constituting 64 percent of the retweets volume of the top 25 tweets. In these 16 tweets, Trump and his son mostly accused mainstream media of being fake news, while many other top tweets were supportive of Trump and critical of mainstream media like user @yoiyakujimin, which happens to be a known bot that is currently suspended from Twitter, having stated: ProPublica – #fakenews & #HateGroup funded by @OpenSociety Main presstitutes […].” In response to the popularity of this particular automated post, ProPublica tweeted: “People also buy Twitter bots to harass journalists. We know because it happened to us” (see Angwin, 2017). Other accounts

Rank	Hashtag	Frequency	Rank	Hashtag	Frequency
1.	fakenews	14,258,909	26.	Trumprussia	59,068
2.	MAGA	634,624	27.	veryfakenews	57,685
3.	Trump	470,693	28.	AmericaFirst	56,068
4.	CNN	439,095	29.	MSNBC	54,574
5.	MSM	249,222	30.	DeepState	52,308
6.	QAnon	138,350	31.	FraudNewsCNN	51,470
7.	WeThePeople	135,556	32.	news	51,305
8.	GreatAwakening	123,834	33.	ReleaseTheCures	51,014
9.	FakeNewsMedia	121,284	34.	Treason	50,831
10.	FakeNewsCNN	106,168	35.	SethRich	50,012
11.	obamagate	95,340	36.	TheResistance	46,102
12.	Russia	93,881	37.	Obama	45,715
13.	CNNBlackmail	85,265	38.	FoxNews	43,149
14.	Fakenewsawards	80,650	39.	resist	41,013
15.	Media	77,967	40.	Macron	39,177
16.	AmericanPravda	75,528	41.	Facebook	36,692
17.	POTUS	75,105	42.	nbc	36,624
18.	Propaganda	69,705	43.	pizzagate	36,416
19.	TCOT	69,307	44.	HateGroup	35,854
20.	TrumpTrain	69,292	45.	wapo	35,463
21.	DrainTheSwamp	63,401	46.	antifa	34,837
22.	AlternativeFacts	63,216	47.	FakePresident	34,141
23.	InternetBillofRights	60,757	48.	wednesdaywisdom	33,840
24.	PresidentTrump	60,299	49.	CnnIsIsis	33,307
25.	Democrats	59,395	50.	realnews	33,209

Table II. The top 50 most used hashtags in connection with #fakenews

that are supportive of Trump include @kirstenkellogg_ and @kwilli1046, which both were suspended from Twitter possibly for being bots. Another user @RealAssange that questioned the fact that Hillary Clinton won the popular vote and treated it as fake news also got suspended from Twitter and the account itself is fake masquerading as, or at least leveraging the fame of the founder of Wikileaks (Digital Forensic Research Lab, 2017). As a matter of fact, only top 3 tweets are actually critical of Trump, accusing him or fabricating facts and/or disseminating fake news in order to serve his political agenda (Table III).

To answer the second research question on the Twitter accounts that are the most active in discussing tweets that mention #fakenews, Table II shows that these most active accounts sent a total of 305,364 tweets (average 6,107 tweets per user) referencing #fakenews, where @PropOrNotApp alone sent 23,863 tweets. It is important to note here that the latter account, which scored 1.3 as being a bot, is associated with the non-partisan group of researchers who run the website "Is It Propaganda Or Not?" (www.propornot.com). They describe themselves as follows: "We are an independent team of concerned American citizens with a wide range of backgrounds and expertise, including professional experience in computer science, statistics, public policy, and national security affairs. We are currently volunteering our time and skills to identify propaganda – particularly Russian propaganda – targeting a US audience" (The PropOrNot Team, 2016). In fact, the list of Russian trolls that is provided by this group has been largely contested as some users have proved to be politically independent rather than partisan sides (Timberg, 2016).

In total, there are 18 accounts suspended by Twitter from this analysis as of May 2018 allegedly for violating Twitter's automation rules (https://support.twitter.com/articles/76915) which are related to "abuse[ing] the Twitter API or attempt to circumvent rate limits." Out of the remaining 32 accounts, the majority ($n = 17$) showed clear affiliation with, support for Trump, or conservative groups such as @Free_PressFail ($n = 11{,}676$), @trey_vondinkis ($n = 6{,}926$) and @avonsalez ($n = 19{,}280$) who describes herself as follows: "I wreak havoc on Libtards with victim cards. #Navymom#Deplorables #MAGA #Americafirst #QArmy #PATRIOT." On the other hand, six Twitter users showed support the democrats or were anti-conservative such as @alternatfacts ($n = 9{,}822$) that has Trump as part of his/her Twitter profile picture with the statement: "President of fake news," while @samir0403 ($n = 5{,}928$) describes himself as follows: "I am an Indian. Got active on twitter on Nov 8th 2016. I was amazed how America can elect such a soulless pathetic human @POTUS." Finally, the remaining nine users had either neutral or unclear political affiliations (Table IV).

By using botometer (https://botometer.iuni.iu.edu), an API developed by a team from Indiana University, we investigated the top 32 accounts (see Table II). The algorithm used indicates scores from 0 for being human-like and 5 for performing like a bot, while scores "in the middle of the scale is a signal that [the] classifier is uncertain about the classification" (BotorNot, 2018). For example, @gerfingerpoken, the Twitter account of Sobieski described earlier as a cyborg was determined by the botometer algorithm as having a score of 1.6 of being a bot; hence, a score 3 and above is more likely to be a bot. Accordingly, we found that the average bots' score is actually 2.3 which means that the classifier is generally not certain about the nature of these accounts. However, 12 accounts scored 3 and above with the highest being 4.6 such as @_breitbot_. If we take into consideration the suspended Twitter accounts ($n = 18$), we conclude that the majority of the top Twitter users that disseminated posts referencing #fakenews are bots ($n = 30$), constituting 60 percent of the total.

According to Kollanyi, Howard and Woolley, bots exhibit "a high level of automation as accounts that post at least 50 times a day, meaning 200 or more tweets, [for it] […] is difficult for human users to maintain this rapid pace of social media activity without some level of account automation" (Kollanyi*et al.*, 2016, pp. 2 and 3). In early August 2017, Twitter suspended the account of "Nicole Mincey" who received praise from Donald Trump himself

Rank	Retweets	Frequency
1.	RT @realDonaldTrump: I am extremely pleased to see that @CNN has finally been exposed as #FakeNews and garbage journalism. It's about time!	39,474
2.	RT @realDonaldTrump: Because of #FakeNews my people are not getting the credit they deserve for doing a great job. As seen here, they are ALL doing a GREAT JOB! https://t.co/1ltW2t3rwy	30,158
3.	RT @realDonaldTrump: I am thinking about changing the name #FakeNews CNN to #FraudNewsCNN!	29,640
4.	RT @MarkRuffalo: Every day it becomes clearer and clearer. The reason @realDonaldTrump labeled legit news #FakeNews early on was because he knew one day all of his deceit, cheating, and harassment, would come under scrutiny by them. The truth has always been his enemy and he knew it	29,482
5.	RT @realDonaldTrump: We will fight the #FakeNews with you! https://t.co/zOMiXTeLJq	25,441
6.	RT @realDonaldTrump: The #FakeNews MSM doesn't report the great economic news since Election Day. #DOW up 16%. #NASDAQ up 19.5%. Drilling &…	21,496
7.	RT @realDonaldTrump: NBC news is #FakeNews and more dishonest than even CNN. They are a disgrace to good reporting. No wonder their news ra…	20,303
8.	RT @yoiyakujimin: ProPublica - #fakenews & #HateGroup funded by @OpenSociety Main presstitutes: @lkirchner @thejefflarson @JuliaAngwin @i…	18,788
9.	RT @kurteichenwald: Ive checked all of @realDonaldTrump's #fakenews declarations from Nov to March. All of them have since proved true in s…	17,842
10.	RT @realDonaldTrump: ….the 2016 election with interviews speeches and social media. I had to beat #FakeNews and did. We will continue t…	17,216
11.	RT @realDonaldTrump: The @NBCNews story has just been totally refuted by Sec. Tillerson and @VP Pence. It is #FakeNews. They should issue a…	16,985
12.	RT @kirstenkellogg_: ProPublica is alt-left #HateGroup and #FakeNews site funded by Soros. @ProPublica @lkirchner @thejefflarson @JuliaAng…	16,766
13.	RT @realDonaldTrump: .@CNN is #FakeNews. Just reported COS (John Kelly) was opposed to my stance on NFL players disrespecting FLAG ANTHEM …	16,132
14.	RT @markantro: CNN creating the narrative #FakeNews https://t.co/nwxizDhTED	16,086
15.	RT @realDonaldTrump: It is my opinion that many of the leaks coming out of the White House are fabricated lies made up by the #FakeNews med…	16,015
16.	RT @realDonaldTrump: To the people of Puerto Rico: Do not believe the #FakeNews! $\#PRStrong_{PR}$	15,435
17.	RT @kwilli1046: Isn't It Interesting How The #FakeNews Media Can't Get Off "The Stormy Slept With Trump" Story But Somehow Congress Never Provided The List Of "Congressional Sexual Predators" Who Used Tax Payer Money To Hide Their Indiscretions In Office. There's a Story That Needs Resolution!	15,111
18.	RT @realDonaldTrump: One of the most accurate polls last time around. But #FakeNews likes to say we're in the 30's. They are wrong. Some…	14,946
19.	RT @realDonaldTrump: 'BuzzFeed Runs Unverifiable Trump-Russia Claims' #FakeNews https://t.co/d6daCFZHNh	13,574
20.	RT @TeaPainUSA: Perfect example of Russian troll farms coordinatin' with far-right nutball blogs to generate #FakeNews and further Trump's attack on Mueller. Notice they are not RTs, but sent as original content. Yet, each tweet is identical. This is all cranked out by one Russian operator	13,358
21.	RT @realDonaldTrump: Biggest story today between Clapper & Yates is on surveillance. Why doesn't the media report on this? #FakeNews!	13,199
22.	RT @DonaldJTrumpJr: Getting to read a #fakenews book excerpt at the Grammys seems like a great consolation prize for losing the presidency.…	13,068
23.	RT @realDonaldTrump: ….it is very possible that those sources don't exist but are made up by fake news writers. #FakeNews is the enemy!	12,058
24.	RT @MichaelCohen212: I have never been to Prague in my life. #fakenews https://t.co/CMil9Rha3D	11,924
25.	RT @RealAssange: Democrats and the #FakeNews: "But @HillaryClinton won the popular vote!" Fact: 7.2 million votes were cast by dead people	11,572

Table III.
Top 30 most retweeted posts

Rank	@username	Tweets	Bot	Rank	@username	Tweets	Bot
1.	Grasslanddesign[a]	25,692	–	26.	lawriter33	3,339	1.4
2.	propornotapp	23,863	1.3	27.	immoralreport	3,204	1.8
3.	avonsalez	19,280	1.5	28.	sealeney	3,158	1.5
4.	politicalpopcu1	18,628	4.5	29.	israeli101	3,086	0.2
5.	johnnystarling	16,851	3.3	30.	Idesignwis[a]	3,059	–
6.	theproplist	14,611	1.8	31.	_breitbot_	3,047	4.6
7.	Plivecalmer[a]	12,583	–	32.	fake___newz[a]	2,848	–
8.	free_pressfail	11,676	3.9	33.	Rharrisonfries[a]	2,809	–
9.	alternatfacts	9,822	0.6	34.	Hoffmanllisa[a]	2,808	–
10.	Fauxnewslive	9,723	3.8	35.	kianmcian	2,641	2
11.	Portofaye[a]	7,622	–	36.	pinkpinta13	2,615	1.3
12.	Msmexposed[a]	7,043	–	37.	teespringstores	2,608	3.5
13.	trey_vondinkis	6,926	4.1	38.	brrrrokkkk	2,593	2.9
14.	fakenewsnews247	6,112	3.5	39.	hetzbeweis	2,588	1.8
15.	col_connaughton[a]	5,975	–	40.	michellebullet1	2,414	4.6
16.	samir0403	5,928	0.5	41.	poetreeotic	2,408	1.9
17.	milove131	5,427	1.3	42.	yerissa_blondee[a]	2,384	–
18.	deplorable80210	4,943	3.2	43.	rosenchild	2,378	1
19.	dumptrumpspace	4,490	1.9	44.	rodstryker	2,325	2.4
20.	somuchtest	4,312	3.4	45.	unsolvedrhyme	2,287	3
21.	saul42[a]	3,664	–	46.	turtlewoman777	2,287	1.6
22.	trend_auditor[a]	3,641	–	47.	whiskey999111[a]	2,274	–
23.	Siddonsdan[a]	3,616	–	48.	draco333999[a]	2,271	–
24.	macansharp	3,536	1.4	49.	jedi_pite_bre[a]	2,255	–
25.	Maconnal[a]	3,470	–	50.	solomon99999000[a]	2,244	–

Notes: [a]Account suspended and is no longer available. Data on bots' scores were retrieved from Botometer on May 29, 2018

Table IV.
Summary of the 50 most active Twitter users and their bots' scores

with the username @ProTrump45 for being his super fan. However, Phillip (2017) reported that this account was another bot used to amplify Trump's views and virally disseminate them on the online platform.

In order to further examine the data for the likelihood of being bots, we randomly selected two larger samples of Twitter accounts, each containing 102,000 Twitter users collected between January 3 and July 21 of 2017 and another sample collected between January 1 and May 7 of 2018 by using a Python package provided by botometer. It returns a metric of Bot-likelihood, including both overall score and scores in specific categories such as "content" and "temporal." Since its inception in late 2015, the Botometer API (originally named "Botornot") has undergone a few updates in its presentation and algorithm. In its May 10, 2018 update, the complete automation probability (CAP) is introduced. Compared with other older metrics, the CAP value is better calibrated and reflects a more conservative estimation of the likelihood an account is completed automated (thus a real Bot)[1]. For our purpose, we choose the CAP in this study in order to minimize the likelihood of falsely labeling human accounts as bots.

We chose two data sets from two periods in order to examine whether Twitter's decision to crack down on bot accounts which started in June 2017 has been effective in limiting bots' use in the dissemination of #fakenews (Twitter Public Policy, 2018). Due to user security setting as well as Twitter policy (e.g. some accounts have been suspended and no longer available), the API does not return results for all accounts. Eventually, we obtained CAP of 78,132 accounts for the older data set (mean = 0.078, SD = 0.21), and 93,322 for the newer data set (mean = 0.063, SD = 0.17). Given a majority of accounts are not bots (thus with CAP of or close to 0), both data sets are highly skewed

(kurtosis = 11.97 and 15.2, respectively). Figure 5 shows the density plot of the CAP for both data sets after log transformation. As indicated in the two analyses, the CAP value of the first sample is higher than that of the second sample, indicating that Twitter has actually achieved some success in decreasing, but not ending, bots' use.

Discussion and conclusion

The dissemination of fake news discourses can be regarded as a method for networked spamming opponents for a variety of reasons. Most importantly, fake news propagation – much like propaganda models during the World Wars – serves the interests of some groups that benefit from this mistrust in mainstream media in order to further their political, economic and other agendas. While it can be argued that "democracies depend on an informed public, totalitarian regimes on fake news" (Martinson, 2017), it is important to avoid hypodermic-needle theorizing and to position the effects of fake news appropriately (Groshek and Koc-Michalska, 2017). Along these lines, however, and of particular importance to the study reported here, the use of bots is not only aimed at spreading fake news and enhancing a political party's messaging power, but it is also meant to "hack free speech and to hack public opinion" (Timberg, 2017). This is because fake news itself is considered as a potential public threat to the proper functioning of democratic discourse and decision making, and better understanding it is highly relevant today as fake news can undermine democracy and the public's faith in factual, watchdog news organizations. Most importantly, fake news references have become weaponized tools (Al-Rawi, 2018) used as a part of networked political spamming that functions as a proxy to undermine credibility and weaken the opponents' arguments by the association that is made. Here, social media users, who actively support political figures interested in attacking their opponents online, might be unwittingly acting as political spammers in their supportive dissemination activities, while bots greatly enhance this spamming effort.

As can be seen from the above findings on the top 50 most recurrent mentions and hashtags, there is a clear focus on major liberal news organization especially CNN in the discourse surrounding fake news. While it is not possible to understand the tone of this discourse without detailed content analysis, the most frequent hashtags provide insight into

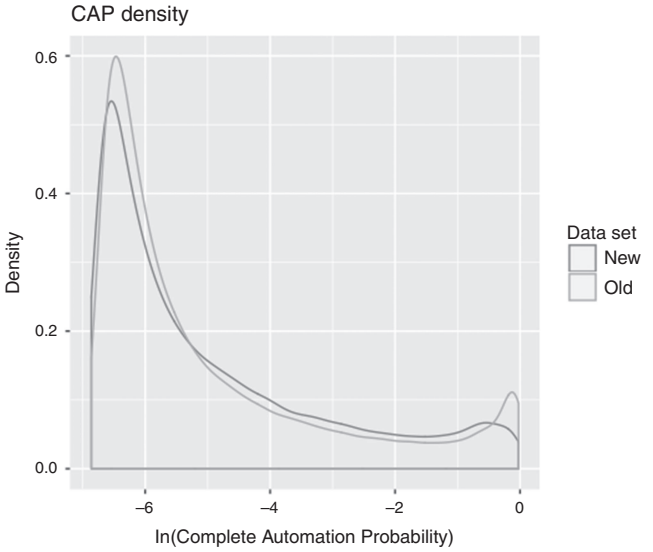


Figure 5.
Bots' scores for two random samples of 204,000 Twitter accounts collected between January 3 to July 21 of 2017 as well as from January 1 to May 7 of 2018

the attitudes associated with CNN and fake news as we can clearly identify the salience of negative terms like #FakeNewsCNN, #CNNBlackmail, #FraudNewsCNN and #CnnIsIsis ($n = 276{,}210$ in total). In fact, there is not a single positive attribute associated with CNN in the most recurrent hashtags. This is also collaborated in the examination of the 25 most retweeted posts as CNN in particular has received the lion's share in the accusations of being a fake news organization mostly due to the popular tweets of Donald Trump and his son, while their supporters have assisted in the propagation of fake news discourses and associating them with other liberal mainstream media like ProPublica and NBC. This shows that conservative groups that are linked to Trump and his administration have dominated the fake news discourses on Twitter due to their activity and use of bots. In fact, top 4 Twitter accounts that showed support for Trump while receiving the highest number of retweets got suspended from Twitter mostly due to being bots which shows the danger of automated accounts that can go viral and move online debates toward certain directions. The above attacks against CNN and other mainstream media outlets would not have been clearly visible without the success of networked political spamming, whether by bots or humans.

Bots aside, there are other subtle yet important mechanisms that ought to be taken into account when addressing the issue of fake news. To begin, although the impact of fake news has often been discussed above within the realm of social media, it is important to note that a part of the disruptive power of fake news lies in its propagandistic and agenda-setting capacity on the entire mediascape that exists beyond social media – and this is reinforced by the fact that a majority of the most active accounts ($n = 30$, 60 percent) including some that belonged to the most retweeted posts ($n = 4$, 16 percent) in this study likely came from spamming bots. Though all social bots are essentially algorithms designed to accomplish simple informational tasks, they are by no means monolithic. Social media platforms are populated by multiple species of bot accounts, employed by entities and organizations with distinctly different agendas. As it relates, a recent study shows that, though not successful on all topics, fake news is especially capable of setting the agenda for key issues regarding international relations, the economy and religion (Vargo *et al.*, 2018). Moreover, in 2016, such an influence is particularly strong on online partisan media, which increasingly serves as an effective conduit to reach legacy news organizations (Vargo and Guo, 2017). In other words, bots have the potential to influence people's agenda especially if the networked spamming messages propagated by these automated accounts go viral such as the case of some of the most retweeted posts examined in this study, for there is a dominant online communication structure that is critical of liberal mainstream media in the way they are mostly associated with references to fake news.

At the same time, some observers see fake news as a problem posing imminent threat to democracy, others working in the area believe that a part of the worry over fake news has been ballooned into a moral panic (Beckett, 2017). As related to his more recent articulation on the public sphere, Habermas (2006, p. 415) wrote: "the public sphere is rooted in *networks* for the wild flows of messages – news, reports, commentaries, talks, scenes and images" (emphasis added). If social media platforms do offer insights into the latest evolution of the structural transformation of the public sphere (Habermas, 1989), they could hardly do so without this unprecedented computational propaganda afforded by networked political spamming.

Fake news remains an important field of study for many contemporary areas of interest. It can instantaneously and easily spread on social media mostly due to its networked affordances and the relative ease of spreading information. Though the peak of fake news stories online was thus far in the period immediately following the US election, other fake reports periodically emerge. Here, it worthwhile to reiterate that many accounts endorsing nearly all political factions and affiliations are responsible for spreading fake news in different levels. As one example, the factchecking website Snopes mentioned that in April 2017 fake

anti-Republican stories started outnumbering fake pro-Republican news stories (BBC Trending, 2017), and it also indicated that many fake stories do not easily cease being shared by people on social media (Criss, 2017), such as the "claim that HIV and AIDS are man-made diseases" (Grimes, 2017). The same applies to the findings of this study as fake news discourses on Twitter seems to be driven by people who belong to all political factions though Trump's supporters remain dominant in the most active users category ($n = 17$, 53 percent). The same finding is observed in the examined top mentions and hashtags which include the names of journalists and politicians from various affiliations and backgrounds.

In sum, this study has provided insight into Twitter users' networked spamming accounts that influenced the discussion on fake news on Twitter. While there is no simple solution to the issue of fake news discourse dissemination, it is all but inevitable that the sophistication and reach of bots and cyborgs will only continue to improve. Our hope is that the reactions of scholars, developers and policy makers can be informed by this contribution that sheds light on why fake news discourses are repeatedly referenced on Twitter and how they are currently used as a weapon to spam opponents. More importantly, the discourses surrounding fake news on social media, which are often amplified by bots, can influence audiences especially in their understanding of what fake and factual news is and their general trust in mainstream media credibility. One of the findings of this study indicates that Twitter has recently succeeded in slightly limiting the use of bots on its platforms, but more efforts are needed to enhance these efforts with broader technical measures. Media educators can make use of this study to further enlighten the public and possibly networked political spammers who are not aware of the nature of their online behavior, by highlighting the potential impact or effect of their online spamming activity.

Of course, there are limitations to this study, including the sampling principally of Twitter and (to the extent possible) future research studies can explore the spread of #fakenews in other platforms like Instagram. Other theories such as selective exposure may also be relevant in understanding the reasons behind the circulation and sharing of fake news by certain users, and determining the effect size of fake news exposure is also critical, and that can be triangulated using big data approaches such as this one with audience surveys and interviews to better explore a still under-researched area of study. Finally, spamming and networked discourses of fake news are limited here to the study of news and politics, but there are other important and vital issues that can be explored using the same or different methodology, since spammers are also active in spreading (dis)information on the environment, health and science.

Note

1. https://botometer.iuni.iu.edu/#!/faq#what-is-cap

References

Al-Rawi, A. (2017a), "Viral news on social media", *Digital Journalism*, pp. 1-17, available at: https://doi.org/10.1080/21670811.2017.1387062

Al-Rawi, A. (2017b), "Audience preferences of news stories on social media", *The Journal of Social Media in Society*, Vol. 6 No. 2, pp. 343-367.

Al-Rawi, A. (2018), "Gatekeeping fake news discourses on mainstream media versus social media", *Social Science Computer Review*.

Angwin, J. (2017), "How journalists fought back against crippling email bombs", *Wired*, September 11, available at: www.wired.com/story/how-journalists-fought-back-against-crippling-email-bombs/ (accessed September 15, 2017).

Bastian, M., Heymann, S. and Jacomy, M. (2009), "Gephi: an open source software for exploring and manipulating networks", *International AAAI Conference on Weblogs and Social Media, San Jose, CA, May 17-20*.

BBC Trending (2017), "The rise of left-wing, Anti-Trump fake news", April 15, available at: www.bbc.com/news/blogs-trending-39592010 (accessed October 20, 2017).

Beckett, C. (2017), " 'Fake news': the best thing that's happened to journalism", *POLIS: Journalism and Society at the LSE*, pp. 1-25, available at: http://eprints.lse.ac.uk/76568/1/blogs.lse.ac.uk-Fake%20news%20the%20best%20thing%20thats%20happened%20to%20journalism.pdf

Bessi, A. and Ferrara, E. (2016), "Social bots distort the 2016 US Presidential election online discussion", *First Monday*, Vol. 21 No. 11.

Bohannon, J. (2017), "Election polling is in trouble: can internet data save it?", *Science*, February 2, available at: www.sciencemag.org/news/2017/02/election-polling-trouble-can-internet-data-save-it

Born, K. and Edgington, N. (2017), "Analysis of philanthropic opportunities to mitigate the disinformation/propaganda problem", Hewlett Foundation, available at: www.hewlett.org/wp-content/uploads/2017/11/Hewlett-Disinformation-Propaganda-Report.pdf (accessed September 17, 2018).

Borra, E. and Rieder, B. (2014), "Programmed method: developing a toolset for capturing and analyzing tweets", *Aslib: Journal of Information Management*, Vol. 66 No. 3, pp. 262-278.

BotorNot (2018), "FAQ", available at: https://botometer.iuni.iu.edu/#!/faq (accessed October 1, 2018).

Cadwalladr, C. (2017), "Robert Mercer: the big data billionaire waging war on mainstream media", *The Guardian*, February 26, available at: www.theguardian.com/politics/2017/feb/26/robert-mercer-breitbart-war-on-media-steve-bannon-donald-trump-nigel-farage (accessed February 27, 2017).

Chu, Z., Gianvecchio, S., Wang, H. and Jajodia, S. (2010), "Who is tweeting on Twitter: human, bot, or cyborg?", *Proceedings of the 26th Annual Computer Security Applications Conference, ACM, December*, pp. 21-30.

Criss, D. (2017), "5 fake stories that just won't go away", CNN, March 10, available at: www.cnn.com/2017/03/10/us/snopes-five-fake-stories-trnd/ (accessed April 19, 2017).

Davis, C., Varol, O., Ferrara, E., Flammini, A. and Menczer, F. (2016), "BotorNot: a system to evaluate social bots", *Proceedings of the 25th International Conference Companion on World Wide Web, Montreal, April 11-15*.

Digital Forensic Research Lab (2017), "Fake assanges drive far-right messages", The Medium, September 5, available at: https://medium.com/dfrlab/fake-assanges-drive-far-right-messages-604a8658bde8 (accessed May 13, 2018).

Ferrara, E. (2017), "Disinformation and social bot operations in the run up to the 2017 French presidential election", *First Monday*, Vol. 22 No. 8.

Ferrara, E. (2018), "Measuring social spam and the effect of bots on information diffusion in social media", *Complex Spreading Phenomena in Social Systems*, Springer, Cham, pp. 229-255.

Forelle, M., Howard, P., Monroy-Hernández, A. and Savage, S. (2015), "Political bots and the manipulation of public opinion in Venezuela".

Frankel, L. and Hillygus, D. (2014), "Niche communication in political campaigns", in Kenski, K. and Jamieson, K.H. (Eds), *The Oxford Handbook of Political Communication*, Oxford University Press, Oxford, pp. 179-194.

Gallacher, J., Kaminska, M., Kollanyi, B., Yasseri, T. and Howard, P.N. (2017), "Social media and news sources during the 2017 UK General Election", available at: comprop.oii.ox.ac.uk (accessed August 25, 2018).

Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y. and Zhao, B.Y. (2010), "Detecting and characterizing social spam campaigns", *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, ACM, November*, pp. 35-47.

Gilani, Z., Crowcroft, J., Farahbakhsh, R. and Tyson, G. (2017), "The implications of twitterbot generated data traffic on networked systems", *Proceedings of the SIGCOMM Posters and Demos, ACM, August*, pp. 51-53.

Grimes, D. (2017), "Russian fake news is not new: Soviet Aids propaganda cost countless lives", *The Guardian*, June 14, available at: www.theguardian.com/science/blog/2017/jun/14/russian-fake-news-is-not-new-soviet-aids-propaganda-cost-countless-lives (accessed May 10, 2018).

Groshek, J. (2014), "Twitter collection and analysis toolkit (TCAT) at Boston University", available at: www.bu.edu/com/bu-tcat/ (accessed August 25, 2018).

Groshek, J. and Koc-Michalska, K. (2017), "Helping populism win? Social media use, filter bubbles, and support for populist presidential candidates in the 2016 US election campaign", *Information, Communication & Society*, Vol. 20 No. 9, pp. 1389-1407.

Grossman, S. (2004), "Keeping unwanted donkeys and elephants out of your inbox: the case for regulating political spam", *Berkeley Technology Law Journal*, Vol. 19, pp. 1533-1575.

Habermas, J. (1989), *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*, Polity, Cambridge.

Habermas, J. (2006), "Political communication in media society: does democracy still enjoy an epistemic dimension? The impact of normative theory on empirical research", *Communication Theory*, Vol. 16 No. 4, pp. 411-426.

Himelboim, I., McCreery, S. and Smith, M. (2013), "Birds of a feather tweet together: integrating network and content analyses to examine cross-ideology exposure on Twitter", *Journal of Computer-Mediated Communication*, Vol. 18 No. 2, pp. 154-174.

Howard, P. and Kollanyi, B. (2016), "Bots, #strongerin, and #Brexit: computational propaganda during the UK-EU Referendum", available at: http://arxiv.org/abs/1606.06356 (accessed August 25, 2018).

Jackson, D. (2017), "Issue brief: distinguishing disinformation from propaganda, misinformation, and 'fake news' ", National Endowment for Democracy, October 17, available at: www.ned.org/issue-brief-distinguishing-disinformation-from-propaganda-misinformation-and-fake-news/ (accessed September 17, 2018).

Just, M.R., Crigler, A.N., Metaxas, P.T. and Mustafaraj, E. (2012), " 'It's Trending on Twitter'-an analysis of the Twitter Manipulations in the Massachusetts 2010 special senate election", APSA 2012 Annual Meeting Paper, pp. 1-23, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2108272

Kollanyi, B., Howard, P.N. and Woolley, S.C. (2016), "Bots and automation over Twitter during the first US Presidential debate", COMPROP Data Memo, available at: https://assets.documentcloud.org/documents/3144967/Trump-Clinton-Bots-Data.pdf (accessed May 15, 2018).

Krueger, B. (2006), "A comparison of conventional and Internet political mobilization", *American Politics Research*, Vol. 34 No. 6, pp. 759-776.

Krueger, B. (2010), "Opt in or tune out: email mobilization and political participation", *International Journal of E-Politics*, Vol. 1 No. 4, pp. 55-76.

Martinson, J. (2017), "A question for a dystopian age: what counts as fake news?", *The Guardian*, available at: www.theguardian.com/media/2017/jun/18/aquestionforadystopianagewhatcountsasfakenews?CMP=Share_iOSApp_Other (accessed May 10, 2018).

Metaxas, P. and Mustafaraj, E. (2009), "The battle for the 2008 US congressional elections on the web".

Musgrave, S. (2017), "Trump address Twitter numbers appear to be boosted by 'bots' ", Politico, January 3, available at: www.politico.com/story/2017/03/trump-speech-twitter-235590 (accessed May 1, 2018).

Mustafaraj, E. and Metaxas, P. (2010), "From obscurity to prominence in minutes: political speech and real-time search", *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, Raleigh, NC, April 26-27*.

Newman, B. (1999), *The Mass Marketing of Politics: Democracy in an Age of Manufactured Images*, Sage Publications, CA.

Phillip, A. (2017), "The curious case of 'Nicole Mincey,' the Trump fan who may actually be a bot", *The Washington Post*, August 7, available at: www.washingtonpost.com/politics/the-curious-case-of-nicole-mincey-the-trump-fan-who-may-actually-be-a-russian-bot/2017/08/07/7aa67410-7b96-11e7-9026-4a0a64977c92_story.html?utm_term=.aad28a95288f (accessed May 23, 2018).

Quinn, P. and Kivijarv, L. (2005), "US political media buying 2004", *International Journal of Advertising*, Vol. 24 No. 1, pp. 131-140.

Raynauld, V. and Greenberg, J. (2014), "Tweet, click, vote: Twitter and the 2010 Ottawa municipal election", *Journal of Information Technology & Politics*, Vol. 11 No. 4, pp. 412-434.

Roller, E. (2016), "The women who like Donald Trump", *The New York Times*, May 10, available at: www.nytimes.com/2016/05/10/opinion/campaign-stops/the-women-who-like-donald-trump.html (accessed May 22, 2018).

Rooksby, E. (2007), "The ethical status of non-commercial spam", *Ethics and Information Technology*, Vol. 9 No. 2, pp. 141-152.

Shao, C., Ciampaglia, G.L., Varol, O., Flammini, A. and Menczer, F. (2017), "The spread of fake news by social bots", July 24, pp. 1-16, available at: https://andyblackassociates.co.uk/wp-content/uploads/2015/06/fakenewsbots.pdf

Siddiqui, S. (2018), "Donald Trump faces backlash as he reveals 'Fake News Awards' winners", *The Guardian*, January 18, available at: www.theguardian.com/us-news/2018/jan/17/trump-fake-news-awards-winners (accessed May 20, 2018).

Sridharan, V., Shankar, V. and Gupta, M. (2012), "Twitter games: how successful spammers pick targets", *Proceedings of the 28th Annual Computer Security Applications Conference, ACM, December*, pp. 389-398.

Sweet, M. (2003), "Political E-Mail: protected Speech or Unwelcome Spam?", *Duke Law & Technology Review*, Vol. 1 No. 1, pp. 1-9.

The PropOrNot Team (2016), "Black Friday Report: On Russian Propaganda Network Mapping", November 26, available at https://drive.google.com/file/d/0Byj_1ybuSGp_NmYtRF95VTJTeUk/view (accessed April 11, 2018).

Timberg, C. (2016), "Russian propaganda effort helped spread 'fake news' during election, experts say", *Washington Post*, November 24, available at: www.washingtonpost.com/business/economy/russian-propaganda-effort-helped-spread-fake-news-during-election-experts-say/2016/11/24/793903b6-8a40-4ca9-b712-716af66098fe_story.html?utm_term=.4fe0be44cf9b (accessed January 11, 2017).

Timberg, C. (2017), "As a conservative Twitter user sleeps, his account is hard at work", *The Washington Post*, February 5, available at: www.washingtonpost.com/business/economy/as-a-conservative-twitter-user-sleeps-his-account-is-hard-at-work/2017/02/05/18d5a532-df31-11e6-918c-99ede3c8cafa_story.html?utm_term=.6f32697a59e3 (accessed March 16, 2017).

Treré, E. (2016), "The dark side of digital politics: understanding the algorithmic manufacturing of consent and the hindering of online dissidence", *IDS Bulletin*, Vol. 47 No. 1, pp. 127-138.

Twitter Public Policy (2018), "Update on Twitter's Review of the 2016 U.S. Election", January 19, available at: https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html (accessed February 20, 2018).

Vargo, C.J. and Guo, L. (2017), "Networks, big data, and intermedia agenda setting: an analysis of traditional, Partisan, and emerging online US news", *Journalism & Mass Communication Quarterly*, Vol. 94 No. 4, pp. 1031-1055.

Vargo, C.J., Guo, L. and Amazeen, M.A. (2018), "The agenda-setting power of fake news: a big data analysis of the online media landscape from 2014 to 2016", *New Media & Society*, Vol. 20 No. 5, pp. 2028-2049.

Verkamp, J.P. and Gupta, M. (2013), "Five incidents, one theme: twitter spam as a weapon to drown voices of protest", FOCI, Washington, DC, August.

Wilkinson, D. and Thelwall, M. (2012), "Trending Twitter topics in English: an international comparison", *Journal of the Association for Information Science and Technology*, Vol. 63 No. 8, pp. 1631-1646.

Wojcik, S., Messing, S., Smith, A., Rainie, E. and Hitlin, P. (2018), "Bots in the Twittersphere. Pew research center: internet & technology", April 9, available at: www.pewinternet.org/2018/04/09/bots-in-the-twittersphere/ (accessed June 23, 2018).

Further reading

Cadwalladr, C. (2016), "Google, democracy and the truth about internet Search", *The Guardian*, December 4, available at: www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook (accessed January 20, 2017).

Ferrara, E., Varol, O., Davis, C., Menczer, F. and Flammini, A. (2016), "The rise of social bots", *Communications of the ACM*, Vol. 59 No. 7, pp. 96-104.

Jun, Y., Meng, R. and Johar, G.V. (2017), "Perceived social presence reduces fact-checking", *Proceedings of the National Academy of Sciences*.

Nguyen, T.T., Hui, P.M., Harper, F.M., Terveen, L. and Konstan, J.A. (2014), "Exploring the filter bubble: the effect of using recommender systems on content diversity", *Proceedings of the 23rd International Conference on World Wide Web, ACM*, *April*, pp. 677-686.

Qiu, X., Oliveira, D.F., Shirazi, A.S., Flammini, A. and Menczer, F. (2017), "Limited individual attention and online virality of low-quality information", *Nature Human Behaviour*, Vol. 1 No. 7, p. 0132.

Rampton, S. and Stauber, J. (2003), *Weapons of Mass Deception: The Uses of Propaganda in Bush's war on Iraq*, Penguin, London.

Sundar, S. (2016), "Why do we fall for fake news?", The Conversation, available at: https://theconversation.com/why-do-we-fall-for-fake-news-69829 (accessed May 10, 2017).

van der Linden, S., Leiserowitz, A., Rosenthal, S. and Maibach, E. (2017), "Inoculating the public against misinformation about climate change", *Global Challenges*, Vol. 1 No. 2, pp. 1-7.

Varol, O., Ferrara, E., Davis, C., Menczer, F. and Flammini, A. (2017), "Online human-bot interactions: detection, estimation, and characterization", *International AAAI Conference on Web and Social Media*.

Corresponding author
Ahmed Al-Rawi can be contacted at: aalrawi@sfu.ca

A corpus of debunked and verified user-generated videos

Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos and Ioannis Kompatsiaris
Centre for Research and Technology Hellas, Information Technologies Institute, Thermi, Greece

Received 20 March 2018
Revised 17 July 2018
2 October 2018
Accepted 4 October 2018

Abstract

Purpose – As user-generated content (UGC) is entering the news cycle alongside content captured by news professionals, it is important to detect misleading content as early as possible and avoid disseminating it. The purpose of this paper is to present an annotated dataset of 380 user-generated videos (UGVs), 200 debunked and 180 verified, along with 5,195 near-duplicate reposted versions of them, and a set of automatic verification experiments aimed to serve as a baseline for future comparisons.
Design/methodology/approach – The dataset was formed using a systematic process combining text search and near-duplicate video retrieval, followed by manual annotation using a set of journalism-inspired guidelines. Following the formation of the dataset, the automatic verification step was carried out using machine learning over a set of well-established features.
Findings – Analysis of the dataset shows distinctive patterns in the spread of verified vs debunked videos, and the application of state-of-the-art machine learning models shows that the dataset poses a particularly challenging problem to automatic methods.
Research limitations/implications – Practical limitations constrained the current collection to three platforms: YouTube, Facebook and Twitter. Furthermore, there exists a wealth of information that can be drawn from the dataset analysis, which goes beyond the constraints of a single paper. Extension to other platforms and further analysis will be the object of subsequent research.
Practical implications – The dataset analysis indicates directions for future automatic video verification algorithms, and the dataset itself provides a challenging benchmark.
Social implications – Having a carefully collected and labelled dataset of debunked and verified videos is an important resource both for developing effective disinformation-countering tools and for supporting media literacy activities.
Originality/value – Besides its importance as a unique benchmark for research in automatic verification, the analysis also allows a glimpse into the dissemination patterns of UGC, and possible telltale differences between fake and real content.

Keywords Video verification, Fake news, Disinformation detection, User-generated content, Social media, Dataset

Paper type Research paper

1. Introduction

User-generated content (UGC), i.e. media content generated by non-professional bystanders during unfolding newsworthy events, has become an essential component of evolving news stories. The ubiquity of capturing devices means that it is very likely that bystanders may be capturing relevant content and sharing it through various web and social media platforms. News professionals are pressed by competition to integrate such content in their stories, but verifying it first is essential to any news provider's reputation (Hermida and Thurman, 2008). Automatic and semi-automatic tools have the potential of considerably easing and speeding up the verification of UGC.

News content verification through automated means is a relatively young field, comprising a set of distinct disciplines, including rumour analysis (Zubiaga *et al.*, 2018), multimedia forensics (Zampoglou *et al.*, 2017), classification of social media content

This work has been supported by the InVID project, partially funded by the European Commission under Contract No. H2020-687786.

This paper forms part of a special section "Social media mining for journalism".

Online Information Review
Vol. 43 No. 1, 2019
pp. 72-88

1468-4527
DOI 10.1108/OIR-03-2018-0101

(Castillo *et al.*, 2011), web mining and multimedia retrieval (Xie *et al.*, 2011). A recent survey (Kumar and Shah, 2018) presented an analysis of known patterns of disinformation dissemination and approaches on the automatic detection of false information.

Datasets are an important asset for understanding and addressing the problem of news content verification, and range from collections of tampered multimedia content and social media posts, to "rumours", i.e. cascades of unverified information. Carefully designed datasets may contribute both to better understanding the patterns of disinformation dissemination and to training and evaluating automatic detection systems.

This paper deals with user-generated video (UGV) verification, specifically with the effort to discern whether a suspect video conveys factual information or disinformation -in other words, for the sake of brevity, if the video is "real" or "fake". The paper presents the first large-scale video verification dataset, consisting of 380 videos and their 5,195 near-duplicates collected from YouTube (YT), Facebook (FB) and Twitter (TW), including a number of fake and real UGVs and numerous other versions of those videos that were consecutively posted online. The dataset is supplemented with 77,258 tweets that contain links to the dataset's videos. The dataset, named Fake Video Corpus 2018 (FVC-2018), which has been made publicly available[1], was gathered using a systematic process and can provide insights into the nature of disinformation, and the types of fake and real content circulating the web. It is also aimed to serve as a benchmark for automatic content verification methods.

2. Related work

The area of multimedia verification consists of several fields of study, tackling various aspects of the problem from different viewpoints.

2.1 Multimedia forensics

A large part of related research concerns tampering detection and image/video forensics algorithms. Proposed algorithms attempt to detect and localise image modifications, either actively by embedding watermarks in multimedia content and monitoring their integrity (Dadkhah *et al.*, 2014; Botta *et al.*, 2015), or passively by searching for telltale self-repetitions (Zandi *et al.*, 2016; Ferreira *et al.*, 2016) or inconsistencies in the image. Such inconsistencies may appear in the pixel domain or the compressed domain depending on the specific process of tampering. A recent survey and evaluation of such algorithms can be found in Zampoglou *et al.* (2017). Generally, such content-based approaches suffer from a number of issues that often render them inapplicable. One problem is their limited robustness with respect to image transformations. When the images or videos are recompressed or rescaled, as it is often the case with social media uploads, the traces of the tampering tend to disappear (Zampoglou *et al.*, 2017). Another limitation is that such approaches are only relevant in specific cases of disinformation. There are cases where a multimedia item is used to convey false information not by altering its content but by altering its context. One typical such approach is to reuse content from a past event and present it as if it was captured during a current one. Another is to misrepresent the content, e.g. the location where it was taken or the identities of depicted people. In such cases, an approach must be able to evaluate the context of the post (e.g. the profile of the uploader, the linguistic characteristics of the accompanying post or the collective characteristics of all posts sharing the same item) rather than its actual content.

2.2 Automated fact checking

In automated fact checking (Hassan *et al.*, 2015), statements are isolated and their veracity is evaluated using reliable databases providing structured knowledge such as

FreeBase and DBpedia. Such approaches are generally useful for assessing claims pertaining to historical truths rather than unfolding events. Furthermore, the automatic extraction of claims that can be cross-checked with a database is very difficult for audiovisual content. Thus, while it is a promising field, it is not currently suitable for UGV verification.

2.3 Rumour analysis

With the rise of social media, attention shifted to other aspects of verification. Twitter and micro-blogging platforms in general have attracted a lot of attention in the recent past. Several approaches operate at the "event" level, e.g. sets of tweets discussing one event or statement. The task of analysing a collection of social media posts around a claim is commonly referred to as rumour detection, defining rumour as a piece of information that may or may not be true (Zubiaga *et al.*, 2018). In that definition, rumour detection refers to the process of gathering all posts related to a rumour. A classification or verification process can then be used to ascertain whether the rumour is true or not. Work by Castillo *et al.* (2011) was the earliest attempt in this category, presenting an algorithm to classify statements pertaining to events into "truthful" or "untruthful". Recent approaches (Vosoughi *et al.*, 2017) attempt to develop methods for estimating the veracity of rumours by aggregating all posts disseminating them. The typical methodology of such methods is generally the same: a number of features are extracted from the tweet texts, the user profiles and the internal structure of the topic (e.g. retweets), and a dataset of annotated rumours is used to train a classifier. A recent survey of approaches and datasets for rumour detection and classification can be found in Zubiaga *et al.* (2018).

2.4 Tweet/post verification

There exist several verification approaches which aim to classify single posts, without taking into account other similar posts. This is an important distinction, since it ultimately concerns the speed at which an investigator can come to a conclusion about a piece of information. From the moment that the first post (tweet) appears making a claim, to the point where enough posts have been gathered into a "rumour", the time delay may be too long for news cycle standards. For that reason, having a system that can operate at the level of single posts is very useful. Such an approach was presented in Gupta *et al.* (2013), where a set of features are extracted from the tweet text and the user profile and are used to classify tweets as truthful or not. A similar attempt (Wu *et al.*, 2015) was used to classify microblog posts from the Sina Weibo platform. The classification of social multimedia by exploiting the associated tweet and user information was the aim of the "Verifying Multimedia Use" benchmarking task, which took place in MediaEval 2015 (Boididou *et al.*, 2015) and 2016 (Boididou *et al.*, 2016). A recent study (Boididou *et al.*, 2018) compares the three top performing methods in this task.

2.5 Contextual video verification

The work presented here is aimed at video verification. With the exception of a body of works in video forensics which, as explained above, have several limitations in terms of applicability, there is one relevant recent work that attempts to tackle the problem using contextual information (Papadopoulou *et al.*, 2017). In this approach, a small dataset of YT videos, called FVC, was used to train and evaluate an automated classifier. The dataset contains around 104 videos annotated as "real" or "fake", and the approach combines a classifier based on video and channel/user metadata features and a second classifier based on comment-based credibility features. This approach is limited in terms of dataset size and its results can only be treated as indicative.

The dataset presented in this paper, called the FVC-2018, builds upon the data of Papadopoulou *et al.* (2017); however, the methodology and scale are significantly different. Besides extending to a much larger number of cases, in this paper multiple copies of the same video are also collected from multiple platforms. This means that the total number of items in the FVC-2018 is an order of magnitude larger than that of FVC, and much more varied. More importantly, it potentially allows us to move beyond single-item methods, to approaches inspired by rumour detection (i.e. exploiting the presence of multiple items in each "case" to be verified to extract features from its collective properties and temporal evolution). This is partly related to the work of Xie *et al.* (2011), where partial duplicates of news-related videos were gathered to analyse the dissemination of so-called "visual memes", i.e. short video segments passed from one uploader to the other. While their work is in some aspects similar to the one presented here, it did not address the problem of fake videos. Furthermore, while Xie *et al.* (2011) track all news-related content regardless of origin, this work deals with UGV verification specifically.

3. Methodology

3.1 Design and concepts

The FVC-2018 is aimed to serve both as a basis for analysis of the dissemination of UGV (real and fake videos), and as an evaluation benchmark for video verification systems. The definition of fake videos used here follows that of Papadopoulou *et al.* (2017) and includes the following:

(1) staged videos where actors perform scripted actions under direction, falsely presented as authentic UGC captured during an event of interest;

(2) videos where contextual information is false (e.g. the claimed video location is wrong);

(3) past videos presented as being captured during unfolding events;

(4) videos of which the visual or audio content has been altered through editing; and

(5) computer-generated imagery (CGI) posing as real.

Real videos are videos that convey actual facts. For the formation of the dataset, this means that they need to have been verified first. Videos of which the veracity could not be confirmed with confidence were not included in the dataset.

There is a limited number of videos that can be collected like this, as the process is constrained to establish the cases of fake and real videos. However, when a newsworthy video is uploaded, and especially when it makes an unusual claim (regardless of its veracity), it tends to get further disseminated by users. That is, people tend to share and re-upload the content, usually with no mention of the original source. These versions are often slightly altered, not only because of the downloading and recompression, but also by various forms of editing, e.g. by adding highlights, slow motion, commentary or by changing the audio. Thus, each newsworthy UGV tends to be followed by a cascade of other versions, and the overall social media activity around them. All this information can be critical for verification, as it can be used both by human investigators and by automatic contextual analysis approaches.

In order to collect alternate versions from videos, search methods combined with near-duplicate detection tools were used. Near-duplicate retrieval is the task of locating (within a given collection), all videos that visually resemble a given query video. While "visual resemblance" is not a strictly defined term, most near-duplicate retrieval tasks focus on the retrieval of different versions of the original content, e.g. following editing,

post-processing, cropping, etc. Near-duplicate video detection algorithms have achieved significant progress in the recent past (Kordopatis-Zilos *et al.*, 2017) and are mature enough for real-world application.

3.2 Dataset collection

The FVC dataset was used as the initial basis of this work. The FVC contains 104 videos, of which 55 were annotated as fake and 49 as real. The authors of Papadopoulou *et al.* (2017) created the dataset over an extended period of time (2016–2017) in cooperation with media experts from the InVID[2] project to serve as a representative collection of past fake videos, and was manually extended to also contain a number of real news-related UGVs. The first step was to extend it with more cases, both fake and real ones. Between April and July 2017, the dataset was manually extended and reached 117 fake videos and 110 real videos. However, manually gathering news-related UGVs is not a straightforward task. In order to further extend the dataset, one additional valuable source was the context aggregation and analysis service[3], which was developed within the InVID project as a tool for video verification. The service, being one of the few publicly available tools for video metadata analysis, generally attracts traffic from verification experts who submit suspicious videos for verification. All videos submitted to the service between November 2017 and January 2018 resulted in an initial pool of approximately 1,600 videos. This set was filtered to remove non-UGC and other irrelevant content, and consecutively, was annotated as real or fake. For this initial annotation, debunking sites such as snopes.com were used in addition to – especially for real content – the general consensus that reliable news sources publish factual content. Furthermore, snopes.com and other debunking sites were consulted in order to collect more debunked fake videos.

Using the above process, an initial set of 380 videos was formed, of which 200 were annotated as fake and 180 as real. Figure 1 presents some indicative cases. This collection is in itself a useful dataset for the field and the result of a significant amount of manual effort and accumulated knowledge. However, a major contribution of this work is the decision to move beyond isolated videos and try to explore how such content is disseminated and reposted through time in various platforms.

Notes: Top: four real videos – a Greek army helicopter crashing into the sea in front of beach; US Airways Flight 1549 ditched in the Hudson River; a group of musicians playing in an Istanbul park while bombs explode outside the stadium behind them; a giant alligator crossing a Florida golf course. Bottom: four fake videos – "A man taking a selfie with a tornado" – CGI; "The artist Banksy caught in action" – staged; "Muslims destroying a Christmas tree in Italy" – out of context, there is no indication that the men are Muslim; "Bomb attack on Brussels airport" – out of context, the footage is from Moscow Domodedovo airport a few years back

Figure 1.
Indicative cases of real and fake user-generated videos

To this end, the following approach was followed in order to end up with a collection of videos that would be highly likely to contain several versions of the query video:

(1) For each video in the original set, extract the video title.

(2) Reformulate the title of the video in a more general form (called the "event title"). For example, a video with title "Video Tornado IRMA en Florida EEUU Video impactante" was assigned the event title "Tornado IRMA at Florida".

(3) Translate the event title from English into four major languages[4]: Russian, Arabic, French and German using Google Translate.

(4) Use the video title, event title and the four translations as separate queries to the three target platforms: YT, Facebook and Twitter. Group all returned videos in a common pool.

(5) Use the near-duplicate retrieval algorithm of Kordopatis-Zilos *et al.* (2017) to search within this pool for near-duplicates of the video.

(6) Apply a manual confirmation step to remove any erroneous results of the method and only retain actual near-duplicates.

Using the above process, the collected videos were organised into "video cascades". The term refers to a set of videos, starting with a first posted video and including all its near-duplicates temporally ordered by publication time.

Methodologically, two further steps were applied to extend and refine the dataset. The first was to submit the URL of the first video of each cascade to Twitter search, and collect all tweets sharing the video as a link. This is a different type of item than the rest of the videos, since it is a link pointing to a video in the cascade accompanied by some text. However, the type of Twitter traffic that a video attracts can be a useful indicator of its credibility. The second was to scan the dataset and separate between "fake" items that reproduce the falsehood from ones that debunk it or use it for entertainment purposes, and correspondingly for "real" items. This is important because such videos should not be taken into account by certain analysis tasks.

3.3 Verification algorithm

One of the potential uses of the FVC-2018 dataset is to train and evaluate automatic verification algorithms. As a set of baseline results, such an algorithm (Papadopoulou *et al.*, 2017) was applied to the dataset, and its performance on the data was evaluated.

The algorithm builds classification models over two sets of features (Table I). The first set is based on the video metadata, and specifically linguistic features extracted from the video description text and statistics extracted from the video channel. These are concatenated to form a feature vector. The second feature set is based on the comments under the video, and the descriptor is extracted using a two-level approach. First, a set of features is extracted from each individual comment, as shown in Table I. Then the credibility of each comment is independently evaluated using a pretrained model (Boididou *et al.*, 2018). While the model of Boididou *et al.* (2018) was trained on a large set of tweets using their linguistic features, it can similarly be applied to video comments. The classifier returns a credibility score in the range [0, 1]. By accumulating the scores of all video comments in a single 10-bin histogram, a vector of 10 variables is produced per video and used for the comment credibility classifier.

In Papadopoulou *et al.* (2017), metadata and comment descriptor vectors were used to train support vector machine classifiers. In the verification methodology used here, the same process was used, with the addition of two model variants: a concatenation of the two feature sets; and the agreement-based approach of Boididou *et al.* (2018) was used, where,

Video metadata features	Comment credibility features
From channel description	01: text length
01: channel view count	02: number of words
02: channel comment count	03–04: contains question/exclamation mark (Boolean)
03: channel subscriber count	05–06: contains happy/sad emoticon (Boolean)
04: channel video count	07–09: contains 1st/2nd/3rd person pronoun (Boolean)
From video description	10: Number of uppercase characters
05: text length	11–12: number of positive/negative sentiment words
06: number of words	13: number of slang words
07–08: contains question/exclamation mark (Boolean)	14–15: has ":" symbol/"please" (Boolean)
09–10: contains 1st/3rd person pronoun (Boolean)	16–17: number of question/exclamation marks
11: number of uppercase characters	18: readability score
12–13: number of positive/negative sentiment words	
14: number of slang words	
15: has ":" symbol (Boolean)	
16–17: number of question/exclamation marks	

Table I.
Overview of video metadata and comment credibility features

following an initial classification of all videos using the two classifiers, the cases where the two classifiers agree are kept. The rest are re-classified using a concatenated feature, either trained on the original training set, or on a new training set consisting of the original plus the agreed-upon videos. Such approaches have been shown to increase classifier performance and were incorporated in the verification results presented with the dataset.

The dataset presented here contains videos from three different platforms. This presents a challenge in the form of unifying the descriptors and treating each video indistinguishably regardless of platform. The main issue with this is the fact that channel description features are not available for Facebook videos. Thus, in all experiments run in this dataset, whenever videos from all platforms are used, such features are not included. This may lead to a degradation of performance. Thus, for comparison, experiments using only the YT videos of the dataset were also run, to evaluate the potential of platform-specific models.

4. Results

4.1 Dataset overview

The initial, manual collection of videos resulted in 200 unique fake videos and 180 real unique videos. While in many cases these initial, manually collected videos were confirmed to be the first version of the video to be posted, there was no way to confirm this in every case prior to the application of the near-duplicate retrieval approach described in Section 4.1. Following that step, the dataset was extended with 3,729 additional fake videos and 2,283 real videos that partly or fully reproduced the submitted video, published on YT, Facebook or Twitter, and covering a period between April 2006 and June 2018. Overall, 172 fake and 148 real videos had at least one near-duplicate in at least one video platform, while for the rest zero near-duplicates were found. One first noteworthy observation is that the number of collected fake videos is much larger than the corresponding number of real videos. This suggests that fake videos tend to be reproduced more, either as repeating acts of disinformation or in videos analysing, parodying or debunking the video. This is in line with the observations of Vosoughi *et al.* (2017), who, while analysing the distribution of Twitter rumours, also observed a higher rate of reproduction for misleading content.

As described in the "Methodology" section, the next step was to manually study the dataset and categorise the near-duplicates based on their intent. This led to five categories of near-duplicates of fake videos and four categories of near-duplicates of real videos.

The categories for near-duplicates of fake videos are: Fake/Fake: those that reproduce the same false claims; Fake/Uncertain: those that express doubts on the veracity of the claim; Fake/Debunk: those that attempt to debunk the original claim; Fake/Parody: those that use the content for fun/entertainment; Fake/Real: those that contain the earlier, original source from which the fake was made. For near-duplicates of real videos, the corresponding categories are: Real/Real: those that reproduce the same factual claims; Real/Uncertain: those that express doubts on the veracity of the claim; Real/Debunk: those that attempt to debunk their claims as false; and Real/Parody: those that use the content for fun/entertainment. A special category concerns videos labelled Real/Private and Fake/Private, which describes Facebook videos that were relevant to the dataset but were published by individual users and thus could not be accessed through the API in order to extract their context. These were left out of the analysis entirely. Table II shows the number of videos that corresponded to each category and each platform. The column labelled "Total" corresponds to all videos, and does not include the twitter posts that share the video, which are counted separately.

The most time-consuming part of the process was forming and annotating the initial set of 380 videos, which was aided by the fact that most of the videos have already been discussed online. Following that, the annotation process of the near-duplicates was generally straightforward and fast, due to the high degree of content repetition between near-duplicates. As a result, after the authors' team collectively concluded on the appropriate tag for each case's initial video, the individual videos could be correctly and swiftly annotated by a single person. To make the task more manageable, the annotation was assigned to two annotators, each of whom was assigned a different part of the videos to annotate. The annotation required in total roughly 20 h for the YT videos and 20 h for the Facebook videos, while the annotation of tweets sharing the videos took roughly 180 h.

While all types of videos were retained in the FVC-2018 dataset for potential future analysis, the ones considered relevant to the analysis presented here are those which retain the same claims as the initial post, i.e. Fake/Fake and Real/Real. For the rest of this work, all observations and analysis concern exclusively these types of video and ignore the rest.

Overall, the scale of the FVC-2018 dataset is comparable to existing datasets for rumour verification. In comparison, the dataset of Gupta *et al.* (2013) contains 16,117 tweets with fake and real images, while the MediaEval 2016 verification corpus (Boididou *et al.*, 2016) contains 15,629 tweets of fake and real images and videos. The dataset of Vosoughi *et al.* (2017) contains 209 rumours with –on average – more than 3,000 tweets each, of which the collection was carried out automatically in order to reach this scale. One important distinction from rumour verification datasets is that the FVC-2018 cascades were assembled from disassociated videos using visual similarity, and not from a network of replies or retweets. This is important since, in platforms such as YT, such relations between items are not available, which makes their collection a challenging problem.

	Fake videos						Real videos				
	YT	FB	TW	Total	TW shares		YT	FB	TW	Total	TW shares
Initial	189	11	0	200	–	Initial	158	22	0	180	–
Fake	1,675	928	113	2,716	44,898	Real	993	901	16	1,910	28,263
Private	–	*467*	–	*467*	–	Private	–	*350*	–	*350*	–
Uncertain	207	122	10	339	3,897	Uncertain	0	1	0	1	30
Debunk	68	19	0	87	170	Debunk	2	0	0	2	0
Parody	43	2	1	46	0	Parody	14	6	0	20	0
Real	22	51	1	74	0						
Total	2,204	1,133	125	3,462	48,965	Total	1,167	930	16	2,113	28,293

Note: Private videos are not included in the totals

Table II. Types of near-duplicate videos collected

4.2 Video and description characteristics

Certain aspects of the accumulated data can prove useful for the analysis of the dataset and are worth highlighting. These concern the videos themselves, their accompanying text (post, video description) and the posting account. In analysing these characteristics, a first approach is to compare the distribution of various features for fake and real videos. In the rest of this section, when comparing feature distributions we present either the mean or the median, depending on whether the variable follows a normal distribution or not. To evaluate the statistical significance of the corresponding differences, we compare the means using Welch's *t*-test or the Mann–Whitney–Wilcoxon test, respectively, and report the associated *p*-values.

One interesting feature is the difference in video durations. Real videos have an average duration of 149 s if we only take the initial videos into account, and 124 s if we include the edited near-duplicates. In contrast, initial fake videos have a much smaller average duration of 92 s ($p < 10^{-3}$), and their near-duplicates have an average duration of 77 s ($p < 10^{-3}$). While this feature itself is not sufficient to classify a video as fake or real, it is interesting to note that fake videos tend to be significantly shorter. Another interesting distinction is the age of the channel/account posting the video. For real videos, the YT channel median age at the time of posting is 811 days prior to the video publication, while for fake videos this value is 425 ($p < 10^{-3}$). For Twitter, these values are 2,325 and 473 days ($p = 10^{-3}$), and for Twitter shares the corresponding values are 1,297 and 1,127, respectively. In the latter case, while the difference does not seem large, given the large sample size it is still statistically significant ($p < 10^{-3}$). No such information was available for Facebook. Overall, fake videos tend to be posted by "younger" YT and Twitter accounts compared to real videos. With respect to the channel/account, it is also interesting to observe that, for real videos, the median YT channel subscriber count is 349 users, while for Twitter the median follower count is 163,325. The latter value is particularly high due to the fact that only a small number (16) of well-established Twitter accounts with many followers were found to have re-uploaded the content as a native twitter video. This is contrasted to the median number of followers of the Twitter accounts which shared the video as a link, which was 333. For the Fake videos, the median follower counts are much lower: 98 ($p < 10^{-3}$) and 2,855 ($p < 10^{-3}$) for YT and Twitter, respectively. For Twitter shares, the median follower count was 297 which, while closer to the one for real videos, is still significantly lower ($p < 10^{-3}$).

Other interesting conclusions stem from linguistic analysis of the text that accompanies the videos (video description or post text depending on the platform). First, language was automatically detected for all posts using the Python langdetect[5] library. For real videos, the relatively most frequent language in the texts is English (YT: 63 per cent, FB: 41 per cent, TW: 75 per cent and TW shares: 62 per cent). For fake videos, the corresponding values are noticeably lower, although still high (YT: 38 per cent, FB: 28 per cent, TW: 43 per cent and TW shares: 52 per cent). A sizeable number of posts/descriptions did not contain enough text for automatic recognition, although that number was generally smaller for real videos (YT: 13 per cent, FB: 48 per cent, TW: 0 per cent and TW shares: 5 per cent) than for fake ones (YT: 28 per cent, FB: 51 per cent, TW: 0 per cent and TW shares: 4 per cent). Other languages encountered in the set included Russian, Spanish, Arabic, German, Catalan, Japanese and Portuguese. With the exception of Russian fake Twitter videos which are strikingly high (28 per cent), these languages appear at a frequency of less than 6 per cent in each category. A number of features were calculated from this text, namely Polarity, Subjectivity, Flesh reading ease (Kincaid *et al.*, 1975) and the Coleman–Liau index (Coleman and Liau, 1975). Polarity and Subjectivity were calculated using the Python TextBlob library[6] while the other two were calculated using the Python textstat[7] library. No noticeable differences were found between fake

and real videos. Despite the common assumption that fake posts have distinctive linguistic qualities, e.g. stronger sentiment and poorer language (Castillo *et al.*, 2011), no such pattern was found in our study.

4.3 Temporal distribution

An important aspect of the FVC-2018 dataset is the temporal distribution of the near-duplicates, and their relative importance in terms of popularity and user attention. To explore this, a timeline was created (Figure 2), showing all near-duplicates per cascade. This shows how the near-duplicates of real and fake videos are distributed in the dataset. Each line corresponds to one original video and its near-duplicates (i.e. a cascade). The horizontal axis corresponds to the time (log-scale) between the posting of the initial version and its near-duplicates. In principle, each dot corresponds to a near-duplicate being posted at that time, and the colour of the dots corresponds to the platform (red: YT; blue: Facebook; green: Twitter; light blue: sharing the original video link as a tweet). The videos are sorted from the most duplicated ones at the top, to the least duplicated ones at the bottom.

The time range of the dataset reaches a maximum at about 10 years between the first posting of a video and its most recent near-duplicate. The difference in the temporal distribution of the fake videos compared to that of real ones is conspicuous. There are relatively few near-duplicates of real videos posted on YT after 10 days from the original post, and the same pattern also holds for Twitter shares. Instead, for fake videos, near-duplicates are posted at a much higher rate for a much longer interval. This discrepancy is also reflected in the fact that the median time difference between the initial video and its near-duplicates is much higher for fake videos than real ones on YT and Facebook. While for real videos the median temporal distance is one and three days, respectively, for fake videos the corresponding values are 62 ($p < 10^{-3}$) and 148 ($p < 10^{-3}$). For Twitter videos, the values are comparable, one and zero days for real and fake videos, respectively, although the difference is still significant ($p = 3 \times 10^{-2}$), but this concerns only a few items. For videos tweeted as links, the median distance is 6 days for real videos and 27 days for fake videos ($p < 10^{-3}$).

Notes: Red: YouTube; Blue: Facebook; Green: Twitter; Light blue: posting on Twitter (as link) of the first video in the cascade

Figure 2.
Temporal distribution of video near-duplicates

4.4 Video categories

Another interesting feature of the collected videos is the category assigned to them by their uploader. YT and Facebook both have a "Category" tag allowing the user to categorise the video. Figure 3 shows the distribution of category tags for fake and real videos on YT and Facebook. Twitter does not offer a corresponding feature and was thus not considered. One may observe clear differences between fake and real videos on both platforms. Real videos on Facebook tend to be categorised as "News" more frequently than fake ones, where the "Entertainment" and "Music" categories are more prominent. Similarly for YT, categories "Video blogging" and "Comedy" are more frequent for fake videos. Taken alone, this distinction is not enough to identify disinformation. However, it may offer a useful verification signal in some cases.

4.5 Video comments

Another noteworthy component of the dataset is the number and distribution of user comments in the videos. Comments can provide a sort of "wisdom of the crowd", which may assist investigators in identifying inconsistencies that expose fake UGVs, or by providing clues that strengthen the claims made by a video. Previous research (Papadopoulou *et al.*, 2017) has also highlighted the importance of user comments for automatic verification. It is, therefore, important to study the comment distribution in FVC-2018. Overall, the dataset contains 491,636 comments on fake YT videos, and 433,139 comments on real ones, 105,814 and 86,326, respectively, for Facebook videos, and 561 and 215 for Twitter videos (in this case, we treat replies as comments). A significant percentage of these, especially for YT, is found in the first video of each cascade (YT: 81 and 69 per cent for fake and real videos; Facebook: 22 and 9 per cent, respectively). Figure 4 presents the cumulative average number of comments over time per video for the three video platforms.

There are some features that stand out. One is the difference in the number of comments between platforms, with YT attracting significantly more than the others. The second is the difference between the number of comments on fake and real videos on YT, with real ones attracting much more comments. A third observation is the steep increase in the number of YT

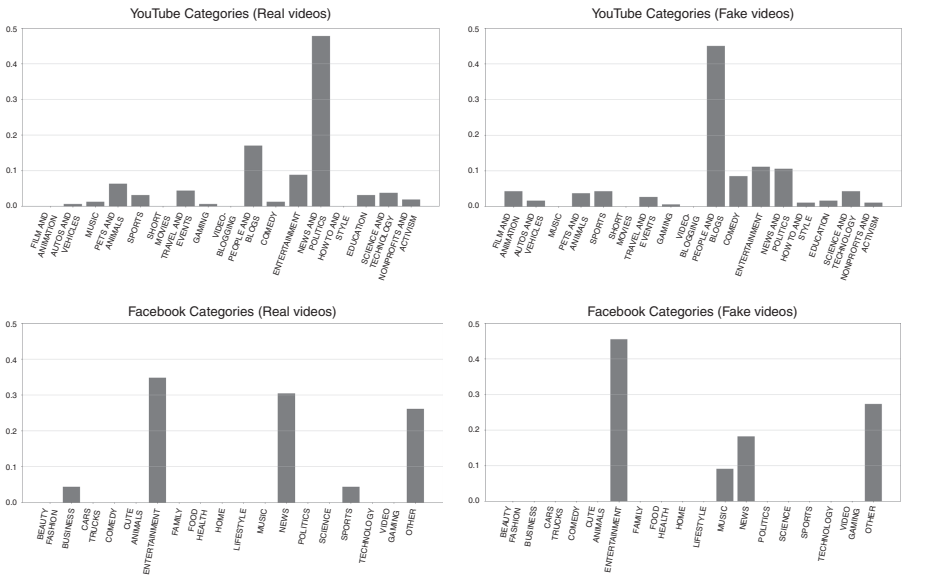


Figure 3.
Video category distribution for YouTube and Facebook

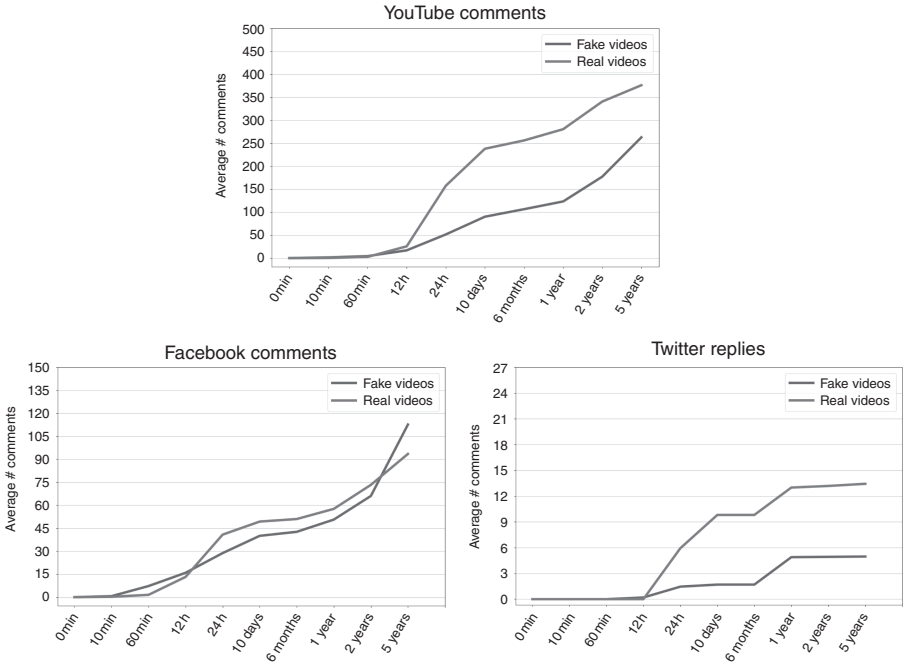


Figure 4.
Timeline view of video comment accumulation

comments in real videos, between 12 h and 10 days after the video is posted, which consecutively tapers off. In contrast, fake videos maintain a steadier rate of accumulation, which, especially after one year from the posting, ends up relatively steeper than for real videos.

4.6 Automatic verification

This section provides a set of experimental results using an existing state-of-the-art automatic verification approach, which can serve as a benchmark for future methods and provides an estimate of the level of challenge that the dataset poses to automatic methods.

The video metadata and comment credibility features described in Section 3.2 were extracted from all videos where this was possible, and the approach of Papadopoulou *et al.* (2017) was applied for verification. Two separate sets of experiments were run: the first set used only the oldest video from each cascade, resembling the setting of Papadopoulou *et al.* (2017). The second set used all videos in the dataset. In both cases, one run considered only YT videos, in order to exploit all possible features, and the other considered videos from all platforms, with reduced features in order to create uniform descriptors. Since there are no first videos from Twitter, the first video per cascade run does not include Twitter videos.

The evaluations were run using 10-fold cross-validation. For the experiments using all near-duplicates, the cross-validation was cascade-based, i.e. all videos from the same cascade were put into the same fold. This ensures that there is no leakage of information between the training and test sets. The results of the runs are presented in Table III.

The rows correspond to different feature sets. The first two correspond to the basic features described in Section 3.2, and the third to their concatenation. "Agreement-based" refers to the practice of separately using the comment credibility and video metadata models, keeping the videos for which the two classifications agree, and re-classifying the rest using the concatenated feature vector. "Agreement-based with retraining" refers to a

Papadopoulou *et al.* (2017)	First video per cascade (YT only)	First video per cascade (YT + FB)	All videos in the cascade (YT only)	All videos in the cascade (YT + FB + TW)
Comment credibility				
Prec.: 0.88	Prec.: 0.91	Prec.: 0.97	Prec.: 0.96	Prec.: 0.94
Rec.: 0.74	Rec.: 0.53	Rec.: 0.52	Rec.: 0.64	Rec.: 0.60
F1: 0.79	F1: 0.67	F1: 0.68	F1: 0.77	F1: 0.73
Video metadata				
Prec.: 0.88	Prec.: 0.87	Prec.: 0.87	Prec.: 0.95	Prec.: 0.95
Rec.: 0.79	Rec.: 0.59	Rec.: 0.58	Rec.: 0.69	Rec.: 0.60
F1: 0.82	F1: 0.70	F1: 0.70	F1: 0.80	F1: 0.74
Concat.				
Prec.: 0.88	Prec.: 0.79	Prec.: 0.77	Prec.: 0.92	Prec.: 0.87
Rec.: 0.82	Rec.: 0.61	Rec.: 0.60	Rec.: 0.70	Rec.: 0.64
F1: 0.85	F1: 0.69	F1: 0.67	F1: 0.79	F1: 0.74
Agreement-based				
Prec.: 0.84	Prec.: 0.58	Prec.: 0.53	Prec.: 0.70	Prec.: 0.61
Rec: 0.88	Rec.: 0.93	Rec.: 0.98	Rec.: 0.96	Rec.: 0.96
F1: 0.86	F1: 0.71	F1: 0.70	F1: 0.80	F1: 0.74
Agreement-based with retraining				
Prec.: 0.77	Prec.: 0.57	Prec.: 0.54	Prec.: 0.69	Prec.: 0.60
Rec.: 0.86	Rec.: 0.92	Rec.: 0.98	Rec.: 0.96	Rec.: 0.97
F1: 0.81	F1: 0.70	F1: 0.69	F1: 0.80	F1: 0.74
Ideal fusion				
Prec.: 1.00	Prec.: 0.64	Prec.: 0.56	Prec.: 0.73	Prec.: 0.64
Rec.: 0.83	Rec.: 0.99	Rec.: 0.99	Rec.: 0.99	Rec.: 0.99
F1: 0.90	F1: 0.79	F1: 0.71	F1: 0.84	F1: 0.78

Notes: In all labels, "Prec": precision, "Rec": recall and "F1": F1-score

Table III. Automatic verification results for the dataset of Papadopoulou *et al.* (2017), the first video per cascade and the entire FVC-2018

similar approach, the difference being that the concatenation-based classifier is retrained using the videos for which the classifiers agreed, thus providing some additional adaptation to the dataset. For every run, the table shows the Precision, Recall and F1-score. Finally, "ideal fusion" is a theoretical result which takes the outputs of the comment credibility and video metadata classifier, and assumes that there exists a perfect fusion system that knows which one is correct on every case. Thus, it correctly classifies a video if at least one of the two results is correct. This provides an estimate of the maximum performance possible if the system had access to the best possible fusion approach.

The evaluation metrics show a degradation of performance on the new dataset compared to the earlier experiments of Papadopoulou *et al.* (2017), both in the case of only using the first posted video in each cascade and when using all the videos in the dataset. When looking at the F1 scores, results are significantly lower than the first column in all cases. Furthermore, removing the channel-based features in order to merge Facebook, Twitter and YT videos leads to significantly reduced performance, both when using only the first video of each cascade and for the entire cascade. Ideal fusion between the two features does increase performance, but not at the levels of Papadopoulou *et al.* (2017).

5. Discussion

5.1 Video and channel characteristics

The aim of this work is to provide insights into the dissemination patterns of fake and real UGVs, in order to assist verification. The results of Section 4 give ground

for several such observations. A first set of observations concern the video and associated text. To the extent that we consider the dataset to be representative of the UGV circulating the web, there are certain patterns that distinguish fake from real videos. One is the length of the video itself, and another is the fact that outlets sharing fake videos tend to be younger and have fewer followers. Another characteristic is the lack of distinctive differences in linguistic terms between fake and real videos. While such features have in the past been used in automatic verification systems (Castillo *et al.*, 2011; Wu *et al.*, 2015; Boididou *et al.*, 2018), it seems that on this dataset these features do not have strong distinctive power. However, that does not mean that such features are useless, since machine learning algorithms may identify complex correlations among them and utilise them for improving classification performance. The same applies to the statistical differences between content categories for fake and real videos on YT and Facebook. While these differences cannot be directly used as a criterion by a human investigator, they may prove useful if combined with other distinguishing features.

Another observation pertains to the large time range that the circulation of a given video may span. In contrast to rumour verification, where generally the duration of rumours is in the scale of hours (Vosoughi *et al.*, 2017), the dissemination of fake videos may cover an entire decade. From a verification perspective, this means that a large bulk of fake videos posted at any given time are actually old fakes that have probably already been debunked. This implies that investigators and the general public could dismiss the majority of misleading content if they had the means to easily match them with their previous, debunked versions.

5.2 Comment distribution

With respect to comment distribution, the overall difference in the average number of comments between fake and real videos can be attributed, at least in part, to the relatively smaller number of near-duplicates for real videos. Indeed, the fake set contains many near-duplicates that attracted relatively little user discussion. On the other hand, real videos are generally not reproduced as much, rather containing a small number of videos with much higher engagement. The high number of near-duplicates (often without comments) for fake videos highlights the fact that only a few fake video reposts have a noticeable disinformation impact.

One explanation for the steady increase of comments in fake videos through time is the tendency of users to link to old fake videos in social media platforms in the context of unfolding news events. This is clearly observed in Figure 3, in which tweets sharing the video links to fake videos appear consistently throughout the decade that the dataset covers. When a video is uploaded with a new date, perhaps slightly edited, it might convincingly appear as a new event and might even mislead an experienced investigator. On the other hand, the practice of simply posting a link to an old fake video is a cruder form of misinformation that is unlikely to fool a professional. It nonetheless seems to be able to misguide part of the public and reinvigorate traffic in an old video, even one that is several years old. Thus, fake videos tend to remain engaging for longer periods, compared to real newsworthy videos which tend to exhaust their engagement with the passing of time both in terms of comment activity in old uploads and with respect to the possibility of being re-uploaded.

5.3 Automatic verification

The observation that user/channel features were important for classification is interesting in the sense that it comes in contrast to the observations of Gupta *et al.* (2013), where user features (on Twitter) led to a degradation of performance. Overall, the experiments imply that existing approaches building on supervised learning and fusion models may be

currently inadequate to deal with the complexity of the problem. This is in contrast to recent works in the related field of rumour/tweet verification, where such methods seem to lead to high classification accuracy (Gupta *et al.*, 2013; Vosoughi *et al.*, 2017; Boididou *et al.*, 2018).

Another important consideration derives from the observation that the experiments using videos from all platforms by removing the channel features led to significant degradation in accuracy. This highlights the importance of platform-specific models for verification. If videos from multiple sources are to be included in an analysis, designing dedicated models for them seems to be necessary in order to achieve good accuracy.

6. Conclusions

This work presented a novel annotated dataset of debunked and verified videos (termed fake and real, respectively) collected from three platforms, and supplemented by the collection of Twitter posts that share the links to them, and organised into cascades. The dataset, named FVC-2018, is also accompanied by a set of experimental results using standard supervised learning and different fusion schemes. Besides its value as a benchmark for future approaches, a third novel contribution of this work is the analysis of the differences in the characteristics and evolution of real and fake video cascades over time with respect to the appearance of near-duplicates, the accumulation of comments and the distribution of various features.

The implications of this work for future research are manifold: the data gathering methodology itself followed a novel protocol, exploiting advances in near-duplicate video retrieval to move from isolated videos into cascades. Furthermore, analysis of the collected data showed certain interesting patterns in real and fake videos, which could be exploited by human investigators and verification algorithms. Also, with respect to automatic verification, while the dataset proved to be challenging for the algorithm used, the increased F1-score for the theoretical ideal fusion classifier over all YT videos shows that there could be potential for a fusion scheme to benefit from the relative complementarity of video metadata and comment credibility features.

The observations made on the distribution of fake videos in particular could also be of immediate practical importance. It was observed that videos keep attracting comments for a very long time after they are posted, as a result of being reposted in social media. In raising awareness against disinformation, flagging such old fakes on the platforms where they have been published could be an easy step that could assist users in identifying them and remove one significant source of disinformation. Also, given that a large proportion of misleading videos posted at any given time are actually near-duplicates of past, debunked videos, identifying them as such would greatly reduce the amount of disinformation circulated. With recent advances in near-duplicate detection, this could be a feasible task. Platforms such as YT already apply near-duplicate detection to detect copyright infringement. If the platforms would be willing to open these functionalities to third parties in order to run their own searches, this could empower investigators and the general public to timely identify and dismiss such fakes.

The current work has certain limitations which leave open possibilities for future research. One is the choice of video sharing platforms. Videos were collected from YT, Facebook and Twitter. Other platforms, such as Instagram, could also be studied in the future. Also, the current work did not take into account another form of information sharing, namely messenger applications such as Viber and WhatsApp. While it is difficult to automatically collect information exchanged through such applications, including cascades from such sources would provide new important insights into the dissemination of disinformation. Another limitation is the mixing of established news outlets sharing real videos, with independent users sharing them. A more refined annotation step could be carried out on the dataset to distinguish between the two.

Furthermore, the limitations of the automatic classification method may be surpassed by taking advantage of the cascade structure provided by the dataset. This would be similar to rumour detection algorithms which exploit not only the individual features of posts, but also the way they are disseminated and distributed within the cascade. While tools such as the InVID plugin (Teyssou *et al.*, 2017) already provide relevant verification functionalities for isolated videos, research in cascade analysis could provide tools with more empowering functionalities, and the dataset presented here is aimed as a suitable evaluation benchmark for such a challenge.

Notes

1. https://mklab.iti.gr/results/fake-video-corpus/
2. www.invid-project.eu
3. http://caa.iti.gr/
4. These languages were selected after preliminary tests indicated that near-duplicate videos appear with increased frequency in these languages.
5. https://pypi.org/project/langdetect/
6. http://textblob.readthedocs.io/en/dev/
7. https://pypi.org/project/textstat/

References

Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O. and Kompatsiaris, Y. (2018), "Detection and visualization of misleading content on Twitter", *International Journal of Multimedia Information Retrieval*, Vol. 7 No. 1, pp. 71-86.

Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D.-T., Boato, G., Riegler, M. and Kompatsiaris, Y. (2015), "Verifying multimedia use at mediaeval 2015", *Working Notes Proceedings of the MediaEval 2015 Workshop, CEUR Workshop Proceedings*, *Wurzen*, *14-15 September*.

Boididou, C., Papadopoulos, S., Dang-Nguyen, D.-T., Boato, G., Riegler, M., Middleton, S.E., Petlund, A. and Kompatsiaris, Y. (2016), "Verifying multimedia use at mediaeval 2016", *Working Notes Proceedings of the MediaEval 2016 Workshop, CEUR Workshop Proceedings*, *Hilversum*, *20-21 October*.

Botta, M., Cavagnino, D. and Pomponiu, V. (2015), "Fragile watermarking using Karhunen–Loève transform: the KLT-F approach", *Soft Computing*, Vol. 19 No. 7, pp. 1905-1919.

Castillo, C., Mendoza, M. and Poblete, B. (2011), "Information credibility on twitter", *Proceedings of the 20th International Conference on World Wide Web, ACM*, *March*, pp. 675-684.

Coleman, M. and Liau, T.L. (1975), "A computer readability formula designed for machine scoring", *Journal of Applied Psychology*, Vol. 60 No. 2, p. 283.

Dadkhah, S., Manaf, A.A., Hori, Y., Hassanien, A.E. and Sadeghi, S. (2014), "An effective SVD-based image tampering detection and self-recovery using active watermarking", *Signal Processing: Image Communication*, Vol. 29 No. 10, pp. 1197-1210.

Ferreira, A., Felipussi, S.C., Alfaro, C., Fonseca, P., Vargas-Munoz, J.E., dos Santos, J.A. and Rocha, A. (2016), "Behavior knowledge space-based fusion for copy–move forgery detection", *IEEE Transactions on Image Processing*, Vol. 25 No. 10, pp. 4729-4742.

Gupta, A., Lamba, H., Kumaraguru, P. and Joshi, A. (2013), "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy", *Proceedings of the 22nd International Conference on World Wide Web, ACM*, *May*, pp. 729-736.

Hassan, N., Adair, B., Hamilton, J., Li, C., Tremayne, M., Yang, J. and Yu, C. (2015), "The quest to automate fact-checking", *Proceedings of the 2015 Computation and Journalism Symposium*, pp. 1-5.

Hermida, A. and Thurman, N. (2008), "A clash of cultures: the integration of user-generated content within professional journalistic frameworks at British newspaper websites", *Journalism Practice*, Vol. 2 No. 3, pp. 343-356.

Kincaid, J.P., Fishburne, R.P. Jr, Rogers, R.L. and Chissom, B.S. (1975), *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*, Institute for Simulation and Training, University of Central Florida, Millington, TN.

Kordopatis-Zilos, G., Papadopoulos, S., Patras, I. and Kompatsiaris, Y. (2017), "Near-duplicate video retrieval with deep metric learning", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 347-356.

Kumar, S. and Shah, N. (2018), "False information on web and social media: a survey", arXiv preprint arXiv:1804.08559.

Papadopoulou, O., Zampoglou, M., Papadopoulos, S. and Kompatsiaris, Y. (2017), "Web video verification using contextual cues", *Proceedings of the 2nd International Workshop on Multimedia Forensics and Security, ACM, June*, pp. 6-10.

Teyssou, D., Leung, J.-M., Apostolidis, E., Apostolidis, K., Papadopoulos, S., Zampoglou, M., Papadopoulou, O. and Mezaris, V. (2017), "The InVID plug-in: web video verification on the browser", *Proceedings of the First Int. Workshop on Multimedia Verification, ACM*, pp. 23-30.

Vosoughi, S., Mohsenvand, M.N. and Roy, D. (2017), "Rumor gauge: predicting the veracity of rumors on twitter", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 11 No. 4, pp. 1-36.

Wu, K., Yang, S. and Zhu, K.Q. (2015), "False rumors detection on Sina Weibo by propagation structures", *31st International Conference on Data Engineering, IEEE, April*, pp. 651-662.

Xie, L., Natsev, A., Kender, J.R., Hill, M. and Smith, J.R. (2011), "Visual memes in social media: tracking real-world news in YouTube videos", *Proceedings of the 19th ACM International Conference on Multimedia, ACM, November*, pp. 53-62.

Zampoglou, M., Papadopoulos, S. and Kompatsiaris, Y. (2017), "Large-scale evaluation of splicing localization algorithms for web images", *Multimedia Tools and Applications*, Vol. 76 No. 4, pp. 4801-4834.

Zandi, M., Mahmoudi-Aznaveh, A. and Talebpour, A. (2016), "Iterative copy-move forgery detection based on a new interest point detector", *IEEE Transactions on Information Forensics and Security*, Vol. 11 No. 11, pp. 2499-2512.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M. and Procter, R. (2018), "Detection and resolution of rumours in social media: a survey", *ACM Computing Surveys*, Vol. 51 No. 2, pp. 1-32.

Further reading

Qazvinian, V., Rosengren, E., Radev, D.R. and Mei, Q. (2011), "Rumor has it: identifying misinformation in microblogs", *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, July*, pp. 1589-1599.

Corresponding author
Markos Zampoglou can be contacted at: markzampoglou@iti.gr

Location impact on source and linguistic features for information credibility of social media

Suliman Aladhadh
Computer Science, School of Science, RMIT University, Melbourne, Australia and College of Computer, IT Department, Qassim University, Qassim, Saudi Arabia, and
Xiuzhen Zhang and Mark Sanderson
RMIT University, Melbourne, Australia

Received 15 March 2018
Revised 5 June 2018
16 September 2018
Accepted 17 September 2018

Abstract

Purpose – Social media platforms provide a source of information about events. However, this information may not be credible, and the distance between an information source and the event may impact on that credibility. Therefore, the purpose of this paper is to address an understanding of the relationship between sources, physical distance from that event and the impact on credibility in social media.
Design/methodology/approach – In this paper, the authors focus on the impact of location on the distribution of content sources (informativeness and source) for different events, and identify the semantic features of the sources and the content of different credibility levels.
Findings – The study found that source location impacts on the number of sources across different events. Location also impacts on the proportion of semantic features in social media content.
Research limitations/implications – This study illustrated the influence of location on credibility in social media. The study provided an overview of the relationship between content types including semantic features, the source and event locations. However, the authors will include the findings of this study to build the credibility model in the future research.
Practical implications – The results of this study provide a new understanding of reasons behind the overestimation problem in current credibility models when applied to different domains: such models need to be trained on data from the same place of event, as that can make the model more stable.
Originality/value – This study investigates several events – including crisis, politics and entertainment – with steady methodology. This gives new insights about the distribution of sources, credibility and other information types within and outside the country of an event. Also, this study used the power of location to find alternative approaches to assess credibility in social media.

Keywords Statistical analysis, Social media, Credibility, Information source, Semantic analysis
Paper type Research paper

Introduction

Social media is an important source of information, particularly about national and international events. Facebook and Twitter users get their primary news from social media (66 and 59 percent, respectively) (Gottfried and Shearer, 2016). In total, 85 percent of Twitter posts have news-based content (Kwak *et al.*, 2010). In crisis situation, such as the Japan earthquakes, Twitter posts give alarms faster than a national new agency (Sakaki *et al.*, 2010).

However, the credibility of social media content about an event is a major concern. Studies show that fake news and rumors spread on Twitter during events, such as the Chile earthquake in 2010 (Mendoza *et al.*, 2010) and Hurricane Sandy in 2012 (Gupta *et al.*, 2013). These rumors spread quickly and are difficult to detect in timely manner thereby having a negative impact on decision making (Oh *et al.*, 2011).

Social media "noisy content" mixes information with high and low credibility. So, assessing information credibility is challenging. While many sources are likely to contribute information about an event, some are more credible than others. Credibility of information can be measured

This paper forms part of a special section "Social media mining for journalism".

Online Information Review
Vol. 43 No. 1, 2019
pp. 89-112

1468-4527
DOI 10.1108/OIR-03-2018-0087

through knowledge about the source. For example, identifiable news sources will be more credible than anonymous sources.

Users in social media research are generally classified based on their location as local or remote. Local users are able to get first-hand information from the event site and are called eyewitness. Remote users share information about the event from a distance. Eyewitnesses are most likely to give first hand information about the event; however, many limitations exist for such accounts (see related work section). Increasing the number of credible sources that can be used to find credible information is essential in order to assess social media information quality.

Knowledge that a source is local or remote can help enhance credibility assessment. First, we can estimate level of credibility of sources in each location. Second, local people are able to understand and interpret the event in terms of geographic cultural and political impacts. Currently, it is common practice for traditional stakeholders (e.g. national press) to contact via social media sources who are close to the place of event to get an update (Dailey and Starbird, 2014). So, information coming from the same region of the event is likely to be richer than remote content.

The language of tweets generated from the same event's region differs from language of tweets from outside that region (Morstatter *et al.*, 2014; Kumar *et al.*, 2014; Cheng *et al.*, 2010). Previous research on social media shows that content of the same credibility level (whether credible or not) shares common practice (Castillo *et al.*, 2013). At the same time previous research present the overestimation of prediction in current credibility models when apply in different domain (Boididou *et al.*, 2014; Aker *et al.*, 2017). However, no study has investigated the effect of location on the behavior of different sources and credibility content.

The research questions that we investigated in this paper are:

RQ1. What are the types of sources expected in different events from both in- and outside the country of events?

RQ2. How linguistic features differ among sources of different type, credibility level, and location?

Related work

In this section, we review the research in the areas of credibility, linguistic, information source and user location in social media. All of these areas are related to this research, and for each one of them we show the research gap in relation to our work.

Microblog credibility

Credibility research in social media has diverse directions based on the methods used. Research has considered tweets' content by choosing a number of content, user and network-based features (Castillo *et al.*, 2013; Gupta *et al.*, 2014). Other research has focused on the content features such as the sentiment of tweets and then used those features to train the model to predict the credibility of a tweet (Mitra *et al.*, 2017b; O'Donovan *et al.*, 2012). Other research has used the source of the tweet, "tweet's author," to assess credibility (Ghosh *et al.*, 2012; Gupta and Kumaraguru, 2012). An overview of the rumor detection and credibility research in social media have been well studied (Zubiaga *et al.*, 2017).

Rumors about an event and their spreading have been studied (Zeng *et al.*, 2016; Aker *et al.*, 2017; Kwon *et al.*, 2013; Qazvinian *et al.*, 2011). Detecting the rumor is achieved by analysis of linguistic features, sentiment and part-of-speech tagger. Recent work shows that credibility prediction models cannot be generalized for different events (Boididou *et al.*, 2014; Aker *et al.*, 2017), as the accuracy of the classifiers drop when apply on event for different domain.

Credibility and linguistic features

Studying the linguistic features related to different credibility level in the web has been studied, studied, Popat *et al.* (2016, 2017) presented an approach to identify true and false textual claim. They studied the linguistic styles of documents related to such claim by using a set of lexicons. It assumes that language of a high credible article is unbiased and objective, while subjective language relate to low credible articles. Also, they include the reliability of the web-source of the article. They found the effectiveness of using the language features with other factors to identify the credibility of a given claim. Horne and Adali (2017) compared between the real and fake news, they found a significant differences between them in structure of title and other language features. Moreover, in large scale study for true and false news distributed between 2006 and 2017 on Twitter, found significant differences in language features of user response to false and true news (Vosoughi *et al.*, 2018). Also, the linguistic features were found to be important factor in identifying the experts in Twitter (Horne *et al.*, 2016). The language of a tweet is influenced by user's location (Cheng *et al.*, 2010; Han *et al.*, 2014), including the linguistic features with other factors like location in social media credibility models can enhance the prediction overestimation in the exist models (Boididou *et al.*, 2014).

Information source

Considering the source of information is a critical part of assessing its credibility. In social media, many researchers have attempted to categorize users into high and low credibility sources to be able to reach credible information. Users post tweets varying from globally well-recognized organizations to locally popular community organizers (De Choudhury *et al.*, 2012), and from specialists in such a domain to fake accounts that steal the identity of other people. Consequently, the quality of information is hugely diverse; finding the users who have highly relevant and credible information about such an event as the source of information is challenging. Methodologies to find authentic sources in social media are different; for example, using the topical content and network structure to rank users based on credibility in a given topic (Canini *et al.*, 2011).

Other research has included user-related data such as tweet content and user profile to build their models in addition to using the social media experts groups. These groups are in the nature of topical expertise directories, such as lists and skills membership on Twitter and Linked-in, respectively (Wagner *et al.*, 2012; Bhattacharya *et al.*, 2014; Ghosh *et al.*, 2012; Bastian *et al.*, 2014). For example, lists in Twitter are an organizational feature created by user to group experts in such topics. Previous research has classified users as high and low based on their topical expertise and local authority and expertise (Cheng *et al.*, 2014). However, all these research were limited by focusing on the information source of a general topic and not for a particular event which may be limited by time and location, such as crisis events. Also, these research limited by their methodologies which rely on geolocation information.

User and event location

Using the location of an event to predict content credibility is an important factor. During events, social media platforms often provide the first alarm: people start sharing information about what is happening. Users from the same event location share specific and accurate information while those further from the event location share general content (Kumar *et al.*, 2013).

Users posting information for a particular event from the same or proximate area of an event are known as "eyewitnesses," they are presumed to have accurate information. Many research has attempted to reach eyewitness authors, such as Diakopoulos *et al.* (2012) and Olteanu *et al.* (2015), but there are many limitations to those studies: very few users who are witnesses to an event share information on social media (Truelove *et al.*, 2014); and witness users are defined via GPS coordinates attached to the tweets; however (~1 percent) of tweets

in Twitter are geo-tagged (Morstatter *et al.*, 2013). To date, there has been no research studying the impact of the distance between user and event locations on the credibility of the information.

Methodology

We followed the next steps to complete this study:

(1) We defined the number of dimensions to classify the social media message at time of events: location, informativeness, source and credibility.

(2) We collected Twitter data for sets of events occurring between 2016 and 2017 using Twitter API. These events were taken from three different topics, entertainment, politics and crisis.

(3) We ran set of crowd sourcing tasks to evaluate and characterize these messages: each event had 1,200 tweets.

(4) We analyzed the interaction between type of author location and his/her messages' characteristics based on the message's informativeness, the used source, credibility and the semantic of the message.

Location dimension

Our research questions have three main components: location, source and credibility of social media content. We categorized the tweets of each event based on physical distance between the source and the event location, and then we used crowd source to annotate tweets' informativeness and sources.

Previous research found that the tweets from the same area of event are different from tweets remote from that area (Morstatter *et al.*, 2014). Moreover, the behavior of users in Twitter differ between countries, mainly they are different of using four main features, hashtags, URLs, mentions and retweets (Poblete *et al.*, 2011).

We examined to the interrelation of user location with the type and quality of messages during different events. Other research was interested to find eyewitnesses to different events, where he/she is the most likely to give the first-hand and credible information about the event. However, in many cases determining the eyewitness is hard and many weaknesses in the exist research (Truelove *et al.*, 2014). Users location can be used to find credible information but not necessarily from eyewitness.

So, we build on the hypothesis that being a local user might increase the credibility of information. Our definition of the author's proximity to the event location is that is within the same country in which the event is occurring, similar to (Kumar *et al.*, 2013), while remote users are those who are outside that country. Then to evaluate the impact of distance to the event, we compare the two categories (local and remote) according to different content types.

We classified tweets of each event into two categories, local and remote tweets. To complete that, there are two ways to determine users' location in Twitter, by using GPS coordinates associated with tweets at the time of post. However, the geo-tagged tweets are < 1 percent (Morstatter *et al.*, 2014). The other way to determine tweet location is through user profile location which is entered by users. In this research we followed the second option as did many other researchers (Sakaki *et al.*, 2010; Mislove *et al.*, 2011; Thomson *et al.*, 2012; Poblete *et al.*, 2011; Kumar *et al.*, 2013). We did not include the automated GPS information because location changes as users are often moving continuously. In this work we are interested in the home country of the user and not in current location, we prefer the user profile location as it more accurately reflects the home country of the user.

Hecht *et al.* (2011) found 66 percent of Twitter's users had a valid geographical location, 16 percent had empty locations and only 18 percent had invalid locations. So, the profile

location field is free text which can include not the geographic location entered by users. To determine a user's home country and then to classify them to one of the two location classes (local and remote), we took the following steps.

Local. For defining a user's location inside the country of event, we used country name in different format, for example, for the event in United kingdom we used all different formats as used in Geonames[1]. Also, we used the large cities of each country of event, a Wikipedia entry was used to define the large cities of each country.

Remote. For users outside the country of event, we used the list of all countries' names in different formats (without names of the country of event). Also, we used the list of top 100 largest cities (excluding the cities of the country of event), where the user is located in one of them to be classified as a remote user. The same methodology has been used by Magdy *et al.* (2016).

Content dimension

To assess the content broadcast during different events we reviewed the previous research that analyzed social media content and credibility, we created set of categories that had been included in other research (Olteanu *et al.*, 2015; Starbird *et al.*, 2010; Vieweg *et al.*, 2010; Gupta and Kumaraguru, 2012; Imran *et al.*, 2013). These categories are bit broader to fit with different events, the large number of messages in this kind of studies, and the limitation of using crowd-source workers to complete annotations.

Informativeness. In this step we assess the status of tweets' informativeness to the event. This process is subjective task since it depends on the person seeking the information, and the context of the event to understand the implications. For this dimension we followed the precedents of (Vieweg *et al.*, 2010; Gupta *et al.*, 2014), to measure how a tweet gives understanding about the situation. The following criteria were used:

- related to the event and informative: when a tweet includes information about the situation and helps a reader to understand what is happening;
- related to the event, but is not informative: when a tweet includes information about the situation but does not help one to understand what is happening;
- not related to the event; and
- not applicable.

Source. When people get an update of an event in social media they look for the source of information, which is a main concern for the quality of the shared information.

The users contributing on social media are diverse and distributed as individuals or organizations, and each category has sub-categories. So we select the sources that are always included in different events; the sources were also included in previous research (Olteanu *et al.*, 2015; Thomson *et al.*, 2012).

For this dimension, we developed a categorization schema based on the following types of sources:

- Government: information published by official (State or local body).
- Businesses: information published by profit-making companies or agencies.
- Non-profit organization: information that publish by non-profit organization.
- Traditional media: information publish by news agencies.
- Journalists: associated with an organization or a freelance journalist.
- Eyewitnesses.
- Politicians.

- Academics, specialists, researchers.
- Digerati.
- Celebrities.
- Outsiders: ordinary or non-identified sources.

Credibility assessment. Assessing the credibility of the information is subjective process, in this research we used the common methodology used by the most of social media credibility research (Castillo *et al.*, 2011, 2013; Gupta *et al.*, 2014). We asked annotators to label the tweets based on the credibility into one of the following categories:

- definitely credible;
- seems incredible;
- definitely incredible; and
- I cannot decide.

The credibility definition was used in this research and was provided to the annotators is "offering reasonable grounds for being believed"[2] and then give an explanation for each category. The criteria for a tweet to be a credible is to be a fact, informative, not a personal point of view (Castillo *et al.*, 2011; Shariff *et al.*, 2014). We noticed "seems not credible" in pilot test a kind of replication with "seems credible" as both of them indicate the tweet is credible and not credible, we only keep one of them to enforce evaluators select from the other options as in Castillo *et al.* (2011), Gupta and Kumaraguru (2012).

Data collection

The goal of this paper is to the informativeness and sources in different locations at the time of an event, and the impact of location on studying credibility in social media. For the purpose of the experiment we collected data of different events having different topics. For each event we identified the most used keywords and hashtags that were used during event window time. We then submitted them to Twitter streaming API[3], to crawl event-related tweets. These tweets were most likely to be representative of the discussion of the event in Twitter. Table I provides information about the users' and tweets' number for each event. We collected events in the same way as was used by Vieweg *et al.* (2010), Thomson *et al.* (2012), De Choudhury *et al.* (2012). In this study we only include the active users who had three or more tweets about the event (most of users only has one tweet), similar to (Vieweg *et al.*, 2010; Starbird and Palen, 2010; Thomson *et al.*, 2012). This threshold for sampling was taken to reduce the noise by capturing active users.

We used crowd sourcing to complete the annotation process of informativeness and source. The crowds' workers were given instructions how to complete each task, including the event name with short description and the link to an outside source to read about the event in more detail. This is further detailed in the following subsections. Also, we gave examples about each event to help them understand the task. We used CrowdFlower platform to complete the tasks[4].

Characteristics of the tasks

For all tasks in this study we included those tweets that were written in English. The crowd's workers were from the same country as the event except that when there were not enough crowds, when we included workers from neighboring countries. We followed this procedure because the local workers are more likely to understand the situational awareness of the event, understand the dialects, locations, entities and the culture of overall place of event. Moreover, as crowd flower platform guidelines, 40 to 50 tweets for each event and task where annotated by the research team. Any worker with less than 80 percent

Event name	Country	Year	Posts	Users	Active users
Apple Event Search keywords: #AppleEvent	USA	2016	1,407,577	771,081	106,067
Summer Olympics Search keywords: #Rio2016	Brazil	2016	3,094,539	1,404,981	236,962
Oscar Academy Award Search keywords: #Oscar #Oscar2017	USA	2017	3,133,296	1,463,300	274,144
Italy Earthquake Search keywords: #ItalyEarthquake #PrayForItaly	Italy	2016	512,798	320,306	30,177
London Attacks Search keywords: #LondonAttacks #Prayforlondon #Londonstrong #Westminster	UK	2017	303,884	212,765	16,564
Cyclone Debbie Search keywords: #debbie #cyclone #CycloneDebbie #tcdebbie Qld Queensland cyclone #BigWet	Australia	2017	89,954	33,129	6,292
Presidential Debate Search keywords: #debatenight	USA	2016	5,443,507	1,819,068	355,895
Presidential Election Search keywords: #ElectionDay #ElectionNight #USElection2016	USA	2016	1,295,766	892,094	77,903

Table I.
The events description used in this research

agreement from the annotated tweets was classified as untrusted. A trusted judgment by a trusted worker took a mean of 8 sec for informativeness task, in the source task took 22.75 sec to be completed, while the trusted judgment in the credibility task took 18 sec. The overall agreement between workers judgments for 100 randomly selected tweets by the platform in the informativeness task is 72.5, 81.2 percent for the source task and 81.24 percent for the credibility assessment. For each tweet in all tasks we collected at least three judgments, and the final label was calculated by the majority.

For each step of annotation which includes 1,200 tweets for a single event in each task (informativeness, source or credibility) was completed by about 15–25 workers, each worker was limited by 300 judgments for each task and cannot exceed this limit as recommended by the crowd sourcing platform.

The first classification task was to define the event-related tweets. Some tweets may contain event keywords but they are unrelated to the event. So, for each event we annotated 600 related tweets from each location type selected randomly. For all events, we labeled tweets until we passed the limit and kept only the related tweets (informativeness and not), and then classified them based on their source and credibility.

Evaluation of the tasks

Subjectivity in the classification process like tweets' content might affect results, especially in large scale studies such as ours. To evaluate the effect of this subjectivity in our results, we followed the methodology used by Olteanu *et al.* (2015). Independently, two coders of the research team labeled 100 tweets selected randomly from all events. They classified the tweets based on informativeness and source type. The coders have a full background about all events, read the tweets as displayed on Twitter, they visited links (if any) within a tweet and the author profile of the tweet.

We apply Cohens' Kappa (k) to measure the inter-agreement between the two coders, the k formulated as:

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

where $Pr(a)$ is the number of times that the coders agree and $Pr(e)$ is the probability that the coders agree by chance (Carletta, 1996). The results are ($K = 0.80$) for the informativeness task and ($k = 0.89$) for source task. Both values indicate substantial and excellent agreement between coders' labels.

We followed that by comparing labels of the tweets that both coders agree with labels provided by workers of crowd source. The result of informativeness is ($k = 0.77$), source is ($k = 0.79$) and ($k = 0.81$) for credibility. The results also show substantial agreement in all tasks. Next, we checked the agreement between each author individually with workers. This includes the labels with no agreement between the two authors. The results also indicate high agreement as well: ($k = 0.78$ and 0.64) for informativeness ($k = 0.79$ and 0.72) for source and ($k = 0.80$ and 0.74) for credibility task.

From the previous experiment we can note that crowd source workers provide a reliable set of collective labels in social media labeling tasks. This conclusion similar to the previous studies which used crowd sourcing for labeling (De Choudhury *et al.*, 2012; Diakopoulos *et al.*, 2012). We received 28,800 labels ($8 \times 1{,}200 \times 3$) for each task informativeness, source and credibility.

Results

In this section we present the analysis that we performed on the data received from the crowd source workers. We first present the distribution of the content across locations. Then we study the proportion of the linguistics features among different sources and credibility levels, and the impact of other factors (location and topic) on the tweets' linguistic distribution.

Content vs location dimensions

We begin by presenting the content distribution for each content type and then the content categories across locations.

Informativeness. The distribution of messages that were found to be related to the event (including the first two categories) was on average 91 percent. The proportion of related messages based on location had a similar ratio, the average of local related messages was 88 percent, while the average of remote messages was 91 percent. The effect of distance on the related messages is therefore weak.

The informative messages (only the first category of the informativeness task was considered) gave an average 46 percent, similar to (Gupta *et al.*, 2014). The effect of distance to event location on informativeness of tweets was high; the percentage of informative remote tweets were higher than local tweets; the average was 43 percent in local location, while it was 49 percent in remote location. This indicates that tweets from outside the country of an event post more informative information than those within that country.

Sources. Table II shows the numbers of sources, and their distribution in local and remote locations from all events. For each source we present the proportion in both location categories, we apply sign test (Gibbons and Chakraborti, 2011) to see whether the distribution of sources are different between the two locations. Also, we show the proportion of each source category across topics:

- Government: 3.4 percent of sources in all events are government, they count 4.2 percent of local and 2.7 percent of remote sources (the remote sources are found in crisis event to support a foreign country, or in "Rio2016 Olympic" supporting their national team), $p < 0.05$. As expected the government sources in the country of event are higher than those in outside. Government accounts in entertainment, politics and crisis were 3.3, 0.3 and 5.7 percent, respectively.

	All	SD_{all}	Local	SD_{local}	Remote	SD_{remote}	$P_{two\ locations}$
Government	330	57.078	201	34.668	129	22.931	*
Non_Profit	258	36.394	158	22.05	100	15.005	
Business	141	13.917	62.0	5.946	79.0	8.543	**
Traditional_Media	2,090	130.001	777	60.523	1,313	78.991	**
Eyewitness	8.00	1.309	4.00	1.069	4.00	1.069	
Journalist	625	69.221	374	46.775	251	22.557	**
Academic	78.0	6.319	38.0	4.496	40.0	3.586	
Politician	256	49.702	154	24.33	102	26.612	*
Digerati	259	26.104	143	14.446	116	12.259	
Celebrity	328	35.21	160	18.189	168	17.542	
Ordinary	5,227	222.014	2,729	122.092	2,498	108.39	*

Notes: *$p < 0.05$; **$p < 0.001$

Table II.
The overall sources, and the sources' distribution across the locations

- Non_profit organization: 2.7 percent of all sources in this study are NGOs, 3.3 percent of them are locally and 2.1 percent are remotely, $p > 0.05$. The distribution of these sources differ between topics: 4.1 percent in entertainment; 1.2 percent in politics and 2.3 percent in crises.
- Business: we found 1.5 percent are business sources, the local business sources are 1.3 percent, while the remote business sources are 1.6 percent, $p < 0.001$. The entertainment events were include in the highest number of business sources 3 percent comparing to the politics 0.7 percent and crisis 1 percent.
- Traditional and internet media: these form the second largest source by 22 percent. The traditional sources used locally is 16 percent compared to 27 percent remotely, $p < 0.001$. The traditional sources in crisis events are 29 percent, entertainment events 22 percent and politics events are 11 percent. The number of traditional sources used by remote users is higher than those used by local people in almost all events: "the Cyclone Debbie" occurred in Australia 2017 was the event with greatest traditional media sources by 44 percent.
- Journalist: 6.5 percent of sources in all events are journalists, with 8 percent locally and 5 percent remotely, $p < 0.001$. The journalists in crisis events were the largest group by 8 percent, followed by entertainment 6 percent and then politics 4 percent. Then number of journalist sources was the highest in two events: the Summer Olympics 2016 recorded 11 percent and Cyclone Debbie 2017 recorded 19 percent.
- Academic and researcher: 1 percent of sources belong into researcher and academic, and their distribution is the same 1 percent locally and remotely, $p > 0.05$. The distribution of this source between the two locations is almost the same in all events in different topics.
- Eyewitness: in our results the eyewitness sources were the least estimated sources with < 1 percent. That is because this type of source is most likely related to crises events with diffused and progressive nature (Olteanu *et al.*, 2015).
- Politician: 3 percent of sources are politician, with local sources 3 percent and remote sources measuring 2 percent, $p < 0.05$. Of course, the largest proportion of this source type will be in political events with 8 percent and the "US Presidential Debate" was the event with the largest political sources by 13 percent of the event sources.
- Digerati: the proportion of digerati source is 3 percent, and their distribution locally and remotely are 3 and 2 percent, respectively, $p > 0.05$. Digerati are those people

with most closely related to technology and blogging 4 percent in both entertainment and politics, while it was only 1 percent for crisis.

- Celebrity: celebrities count 3.4 percent of sources from all included events. 3 percent of local and 4 percent of remote sources are celebrities, $p > 0.05$. As with digerati, celebrities as a source have larger proportion of users: 5 percent on both entertainment and political events of their sources compared to only 2 percent in crisis events.
- Ordinary: among all source categories, the ordinary source provides the largest sources 54 percent. The ordinary source locally is 57 and 52 percent remotely, $p < 0.05$. The results indicate that ordinary individuals are the majority of sources in most of events. This finding concurs with those of (Olteanu *et al.*, 2015; De Choudhury *et al.*, 2012). In time of crisis, 50 percent of sources are ordinary and there is a large difference in the distribution of these sources locally 54 percent and remotely 45 percent, compared to events in entertainment 52 percent and politics 67 percent. However, "Cyclone Debbie" was an event with few ordinary sources: 18 percent. "Cyclone Debbie" was expected for a while before it happened, and this affected substantial part of Australia.

Source vs informativeness. Figure 1 presents the interaction between the sources and the informativeness, for example, Figure 1(a) shows the sources distribution with the first informativeness category (related and informative). While in Figure 1(b), we see the distribution of the source in the second informativeness category (related but not informative). Eyewitness was not included due to small sample size (only 8). We can see that some sources like (government, non-government, traditional media and journalist) were high in the informativeness categories, while other sources were high in the not informative category like (politician, digerati and ordinary). A source like (business) has the same ratio in the both categories.

Credibility. The results of the credibility annotations of the tweets were 44.13 percent (4,236 tweets) "Definitely credible," 54.5 percent (5,229 tweets) "Seems incredible," only 1.34 percent (129 tweets) "Definitely incredible" and < 1 percent (6 tweets) "I can't decide." Next, since "seems incredible and definitely incredible" belong to incredible category, we combined them in one class called "incredible," same to (Castillo *et al.*, 2011). The "I can't decide" tweets were discarded, so we end up with two credibility classes "credible and incredible." There is impact of location on content credibility, 41 percent of local tweets were

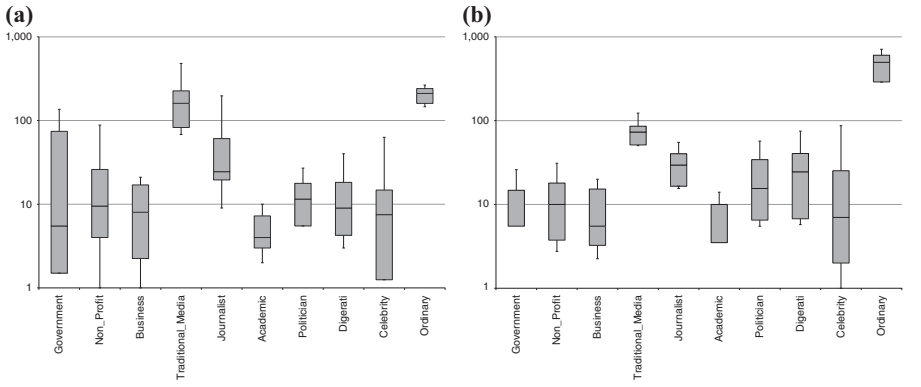


Figure 1.
The informativeness status of each source type

Notes: (a) Related and informative; (b) related but not informative

credible comparing to 48 percent remotely. On the other hand, 59 percent of local tweets were incredible, while it was 52 percent remotely. We can see that the content of the remote sources are more credible than the local ones.

Linguistic feature analysis

Having studied the distribution of the sources and credibility between event, location and topic, we next analyze the features of the tweet content. There are many types of features in Twitter, generally classified into content-based, social-based and network-based features (Castillo *et al.*, 2013; O'Donovan *et al.*, 2012; Kang *et al.*, 2012). Although credibility research in social media has grown rapidly, investigations into the linguistics features around credibility are few.

We used a tweet's language as the indicator of credibility for many reasons. First, the linguistic features of tweets have been found to be important predictor for credibility in social media events (Mitra *et al.*, 2017b; Kwon *et al.*, 2013); second, the language of the tweets is influenced by user's location (Cheng *et al.*, 2010; Han *et al.*, 2014) and that allows us to study impact language and location on credibility; and third the text is available in all tweets whereas other features might be absent.

We used a sentiment analysis tool called Linguistic Inquiry and Word Count (LIWC), which is an application analyzing text by counting words of different psychological categories (Tausczik and Pennebaker, 2010). Its dictionary of categories include almost 6,400 words and word stems. Use of LIWC is common in social media data analyses (Nguyen *et al.*, 2013; Golbeck, 2016), and for credibility research in specific areas such as (Gupta and Kumaraguru, 2012; Kwon *et al.*, 2013; Zeng *et al.*, 2016; Rosso and Cagnina, 2017; Mitra *et al.*, 2017b). The categories we used are listed in Table III and have been used in the previous research. All of these categories except the last one in the table have subcategories, for example, affect feature is the main category, which includes two subcategories: positive and negative emotions. The negative emotion includes three subsub categories anxiety, anger, and sad. So in our analyses, we include general categories; the full list of both general and sub-categories are available at[5], the results were normalized (0-1). We investigated whether any of the categories dominantly appear in such source, credibility level, location and topic.

Source and linguistic features distribution. In this section we present the features across different sources and the source interaction with location and topic. For each factor and factorial interaction we showed The Pillai's trace result, which shows how all of these outcomes variables together differs between group such as source, credibility, location and topic. Then we follow that by investigating the individual Anova for each linguistics outcome.

Source. Since we have eleven different sources, we investigate how they differ from each other. In our analysis we include nine different sources. (W did not include eyewitness and

Feature	Example
Function	it, to, no, very
Affect	happy, cried
Social	mate, talk, they
Cognitive (cogproc)	cause, know, ought
Perceptual (percept)	look, heard, feeling
Biological (bio)	eat, blood, pain
Drives	ally, win, superior
Relativity (relativ)	area, bend, exit
Informal language (informal)	damn, btw, umm
Authentic, Word count(WC), Qmark, Exclam, Hashtags and URL	count each category in a tweet

Table III. The LIWC categories used to perform analyses on tweet content

researcher as there were not enough samples, 8 and 78, respectively). The Pillai's trace was significant ($p < 0.001$). The separate ANOVA showed significant effects with all features ($p < 0.001$), except percept and drives ($p < 0.05$). This significant effect means that different sources have different styles of writing and the features distribution are different in their content.

Source vs location. After we found there were significant differences in all features between sources, we studied the impact of location on a source's feature distribution. The Pillai's trace showed no significant interaction between source and location. Only the separate ANOVA showed "derive" had a significant interaction, see Table IV. This means that the same source type has a similar feature distribution regardless of the source's location.

Source vs topic. The topic factor influences the feature distribution across different topics. We studied how features distribution differs for the same source in different topics. Pillai's trace shows a significant result, $p < 0.001$, there are significant interactions with all features ($p < 0.001$), except in the bio and exclam categories where there were no significant interactions, refer to Table IV.

Source vs location vs topic. The impact of location on feature distribution of source was weak shown in the interaction between source and location. Here we include the third factor "Topic." In considering the interaction between source and location. The feature distribution differs considerably between topics: the three-way interaction shows a significant results in Pillai's trace ($V = 0.09$, $F(256, 9460) = 1.691$, $p < 0.001$). Many features have significant interactions as in Table IV shows. This result indicates that the feature distribution of source tweets across locations change according to different topics.

Credibility and linguistic features distribution. In this section, we study how the used language different between the credibility levels.

Credibility. We analyzed the distribution of features between the two credibility classes, credible and incredible (Table V).

The Pillai's trace's result shows a significant effect of the credibility levels on the linguistics features distribution ($V = 0.184$, $F(16, 9573) = 134.773$, $p < 0.001$). The two credibility levels are different in many features as shown by the ANOVA results, we found significant differences with all features ($p < 0.001$) except "drives and authentic"

	Source vs location		Source vs topic		Source vs location vs topic	
	F	*p*-value	*F*	*p*-value	*F*	*p*-value
Function	0.927	0.493	4.864	0.000**	1.903	0.016*
Affect	0.164	0.995	2.749	0.000**	1.956	0.012*
Social	1.359	0.209	7.105	0.000**	2.633	0.000**
Cogproc	1.033	0.409	3.093	0.000**	1.016	0.436
Percept	0.369	0.937	5.676	0.000**	0.928	0.535
Bio	0.224	0.987	0.610	0.879	0.860	0.616
Drives	2.386	0.014*	2.796	0.000**	2.905	0.000**
Relativ	0.822	0.583	8.779	0.000**	2.193	0.004*
Informal	0.377	0.934	9.015	0.000**	2.952	0.000**
Authentic	0.462	0.883	6.083	0.000**	1.385	0.138
WC	0.672	0.717	4.866	0.000**	3.041	0.000**
Pronoun	1.026	0.413	4.135	0.000**	0.864	0.612
QMark	0.974	0.454	4.930	0.000**	1.363	0.150
Exclam	1.088	0.368	0.640	0.854	0.996	0.457
Hashtags	0.823	0.582	6.298	0.000**	1.494	0.092
URL	0.874	0.538	12.803	0.000**	2.837	0.000**

Notes: *$p < 0.05$; **$p < 0.001$

Table IV.
ANOVA result for source interaction with other factors

	Credibility vs location		Credibility vs topic		Credibility vs location vs topic	
	F	*p*-value	*F*	*p*-value	*F*	*p*-value
Function	0.207	0.649	7.070	0.001*	13.602	0.000**
Affect	1.221	0.269	4.578	0.010*	11.765	0.000**
Social	4.216	0.040*	135.564	0.000**	0.842	0431
Cogproc	0.737	0.391	12.615	0.000**	5.331	0.005*
Percept	2.470	0.116	20.508	0.000**	0.514	0.598
Bio	0.573	0.449	4.879	0.008*	3.822	0.022*
Drives	7.927	0.005*	23.746	0.000**	9.494	0.000**
Relativ	2.763	0.097	105.807	0.000**	7.862	0.000**
Informal	3.429	0.064	34.155	0.000**	8.056	0.000**
Authentic	4.166	0.041*	52.025	0.000**	3.669	0.026*
WC	7.169	0.007*	23.504	0.000**	9.492	0.000**
Pronoun	5.834	0.016*	66.657	0.000**	6.798	0.001*
QMark	0.288	0.591	46.399	0.000**	0.683	0.505
Exclam	6.871	0.009*	2.521	0.008*	1.414	0.243
Hashtags	0.647	0.421	8.981	0.000**	2.375	0.093
URL	7.090	0.008*	28.862	0.000**	13.034	0.000**

Notes: *$p < 0.05$; **$p < 0.001$

Table V.
ANOVA result for credibility interaction with other factors

were not significant. Moreover, occurrences of some features in the incredible class are higher than those in the credible class (Function, Affect, Social, Cogproc, Bio, Informal, Pronoun, Qmark, Exclam, Hashtag ($p < 0.001$). This indicates that credible content is not necessary connected with more features' presence, this confirm the finding made by O'Donovan *et al.* (2012). On the other hand, only a few features (Percept, Relative, WC, URL, $p < 0.001$) occurred in the credible class more than incredible class.

Credibility vs location. In this section, we study the interaction between credibility and location. For example, the tweets of the same credibility class but from different location will share similar features distribution, or the location factor will effect on the features distributions.

The interaction between credibility and location is significant as Pillai's trace has shown ($p < 0.001$). Individual univariate ANOVAs on the outcome variables revealed significant interactions on "social, drives, authentic, WC, pronoun, exclam and url," ($p < 0.05$). Table VI presents where is significant happened between credibility and location for the features found to be affected by the interaction. A bonferroni correction was applied to mitigate the multiple comparison effect.

The credible class has three features were impacted by location, "drive" was significantly higher in local location than remote, while "WC and URL" were higher in remote than the local location. On the other hand, location effect on five features in the incredible class,

	Credible			Incredible		
	Local	Remote	$P_{\text{two locations}}$	Local	Remote	$P_{\text{two locations}}$
Social	0.136	0.132	0.212	0.164	0.150	0.000**
Drives	0.100	0.089	0.000**	0.0.92	0.0.92	0.909
Authentic	0.186	0.174	0.193	0.175	0.192	0.041*
WC	0.548	0.561	0.004*	0.479	0.476	0.394
Pronoun	0.082	0.079	0.437	0.157	0.141	0.000**
Exclam	0.008	0.008	0.893	0.018	0.013	0.000**
URL	0.166	0.194	0.000**	0.125	0.136	0.016*

Notes: *$p < 0.05$; **$p < 0.001$

Table VI.
Credibility vs location

two of them "social and URL" were higher locally and the rest "authentic, pronoun and exclam" were the opposit. In both classes, Url was higher at a remote location than in local one. This indicates the remote authors share more Urls, regardless of their credibility.

Credibility vs topic. The Pillai's trace result show a significant interaction between credibility and topic with the features ($p < 0.001$). The separate univariate ANOVAs show that all features have significant interaction with credibility and topic. Table VII shows that the features found to be important via the ANOVA test are different between credibility classes in different topics. "Crisis" was the most affected topic, then entertainment and politics. There are some features such as "WC and URL" are highly significant at credible class for all topics. Some features such as "Function, Affect, Cogproc, Pronoun, Qmark and Exclam" are high in the incredible class in all topics. Moreover, some features such as "Informal, Social" are high in incredible class in entertainment and crisis, while the opposite applied in the politics topic as it was high in the credible class. On the other hand, some features are different between the two credibility classes in some topics, such as "Bio" is high in incredible class for entertainment and crisis, and "percept" is high at credible class in crisis while no impact in politics.

Credibility vs location vs topic. In this section we study the three-way interaction between credibility, location and topic. Topic has been found to be important factor that can affect the features distribution in social media (Imran and Castillo, 2015; Boididou *et al.*, 2014).

The Pillai's trace results present a significant interaction of the three factors' interaction ($V = 0.014, F(32, 9575) = 4.115, p < 0.001$). This result shows that the features of different topics differ significantly when interacting with credibility and location. Separate univariate ANOVAs on the outcome variables show significant interactions on 11 features (Function, Affect, Drive, Relative, Informal, WC and Url, $p < 0.001$.) and (Cogproc, Bio and Authentic, $p < 0.05$).

Table VIII shows the mean differences between the features in the credible class in different locations and topics. From the three topics, crisis was the most affected topic by location classes: it had nine out 11 features with significant differences between the two locations. Seven of these features were high locally, and only two features were high remotely: "Informal and URL."

	Entertainment			Politics			Crisis		
	Credible	Incredible	$P_{\text{two levels}}$	Credible	Incredible	$P_{\text{two credibility}}$	Credible	Incredible	$P_{\text{two levels}}$
Function	0.275	0.339	0.000**	0.365	0.398	0.000**	0.323	0.354	0.000**
Affect	0.074	0.090	0.000**	0.090	0.102	0.007*	0.076	0.103	0.000**
Social	0.100	0.137	0.000**	0.187	0.156	0.000**	0.116	0.179	0.000**
Cogproc	0.063	0.109	0.000**	0.139	0.151	0.021*	0.083	0.128	0.000**
Percept	0.053	0.054	0.795	0.044	0.046	0.664	0.052	0.029	0.000**
Bio	0.017	0.029	0.000**	0.020	0.022	0.450	0.024	0.035	0.000**
Drives	0.098	0.078	0.000**	0.097	0.101	0.337	0.087	0.099	0.001*
Relativ	0.194	0.131	0.000**	0.136	0.139	0.582	0.230	0.140	0.000**
Informal	0.153	0.166	0.000**	0151	0.137	0.001*	0.142	0.172	0.000**
Authentic	0.126	0.181	0.000**	0.145	0.181	0.002*	0.269	0.193	0.000**
WC	0.551	0.446	0.000**	0.542	0.494	0.000**	0.570	0.493	0.000**
Pronoun	0.050	0.130	0.000**	0.138	0.161	0.000**	0.053	0.155	0.000**
QMark	0.009	0.048	0.000**	0.018	0.028	0.022*	0.006	0.065	0.000**
Exclam	0.010	0.021	0.000**	0.010	0.017	0.000**	0.002	0.008	0.000**
Hashtags	0.123	0.119	0.149	0.118	0.133	0.000**	0.104	0.114	0.001*
URL	0.224	0.140	0.000**	0.129	0.096	0.000**	0.186	0.154	0.000**

Notes: *$p < 0.05$; **$p < 0.001$

Table VII.
Interaction between credibility classes and topics with different features

	Entertainment			Politics			Crisis		
	Local	Remote	$P_{\text{two locations}}$	Local	Remote	$P_{\text{two Locations}}$	Local	Remote	$P_{\text{two Locations}}$
Function	0.278	0.272	0.633	0.366	0.363	0.848	0.349	0.297	0.000**
Affect	0.070	0.078	0.163	0.087	0.093	0.266	0.085	0.066	0.000*
Cogproc	0.060	0.067	0.378	0.137	0.140	0.671	0.095	0.71	0.000*
Bio	0.014	0.021	0.047*	0.022	0.019	0.414	0.023	0.024	0.848
Drives	0.105	0.090	0.007*	0.097	0.098	0.868	0.097	0.077	0.000*
Relativ	0.139	0.158	0.008*	0.136	0.135	0.967	0.239	0.220	0.004*
Informal	0.152	0.155	0.528	0.150	0.152	0.830	0.136	0.148	0.018*
Authentic	0.121	0.131	0535	0.141	0.148	0.657	0.295	0.243	0.000*
WC	0.531	0.571	0.000*	0.546	0.539	0.452	0.567	0.574	0.320
Pronoun	0.045	0.055	0.165	0.136	0.141	0.507	0.066	0.041	0.000*
URL	0.220	0.227	0.439	0.118	0.141	0.008*	0.159	0.213	0.000*

Notes: *$p < 0.05$, **$p < 0.001$

Table VIII.
Credible vs location vs topic

In the credible class, entertainment and politics were the least sensitive compared to crisis. "Bio, Relative and WC" at entertainment, "URL" at politics and "Cogproc, Informal and URL" at Crisis were significantly higher at remote than local locations. Interestingly, for politics and crisis, URL was significantly high at a remote location. This finding shows that "foreign" tweets (i.e from out-side the country of event) always include third party information.

Table IX presents the mean differences of different features in the incredible class for different topics and locations. In this table, we want to see how tweets from different locations classified as incredible share the same or different linguistics features distribution.

Five features have significant differences between local and remote classes in politics comparing to four for entertainment and crisis. All features with significant difference in politics were locally higher than remotely; while the opposite at entertainment significant features except "Function was higher locally." In crisis the "Informal and Pronoun" were locally higher than remotely and the opposite with the other two features "Relative and WC."

In comparison between Tables VIII and IX, we can realize the impact of location on the linguistic features distribution between the two credibility levels for the same topic. For example, there was almost no impact of location on politics at the credible class while in incredible class we found a big influence of location. The same finding in crisis with an

	Entertainment			Politics			Crisis		
	Local	Remote	$P_{\text{two locations}}$	Local	Remote	$P_{\text{two locations}}$	Local	Remote	$P_{\text{two locations}}$
Function	0.354	0.324	0.000**	0.412	0.384	0.015*	0.350	0.359	0.320
Affect	0.094	0.086	0.101	0.110	0.093	0.006*	0.100	0.106	0.233
Cogproc	0.107	0.112	0.430	0.156	0.146	0.209	0.124	0.131	0.322
Bio	0.031	0.026	0.054	0.021	0.024	0.496	0.035	0.036	0.696
Drives	0.073	0.082	0.028*	0.108	0.094	0.011*	0.096	0.102	0.170
Relativ	0.125	0.136	0.050	0.140	0.137	0.732	0.130	0.150	0.001*
Informal	0.158	0.174	0.000**	0.142	0.133	0.125	0.179	0.165	0.002*
Authentic	0.173	0.188	0.222	0.178	0.185	0.654	0.185	0.202	0.200
WC	0.441	0.451	0.114	0.519	0.469	0.000**	0.479	0.507	0.000**
Pronoun	0.130	0.130	0.918	0.179	0.144	0.000**	0.162	0.148	0.017*
URL	0.123	0.158	0.000**	0.103	0.090	0.128	0.149	0.160	0.130

Notes: *$p < 0.05$; **$p < 0.001$

Table IX.
Incredible vs location vs topic

opposite way, it was highly impacted by location in the credible class and we found less impact in the incredible class. The influence of location on entertainment is same in both credibility level with almost different features.

Discussion

Around the world, many events occur daily, and social media has become an important platform where people read and share information about these events (Kwak *et al.*, 2010). These events are combined from many different topics, regardless of event type or the users who contribute in these events, who can be close to or far from the location of the event. Information manipulated during the event varies in credibility and can include inaccurate and false information such as rumors.

In reviewing the literature, no data were found regarding the association between credibility and other factors such as location, topics and linguistic features. Olteanu *et al.* (2015) found that sources and information type differ in different events, and location is found to affect user behavior in terms of language use when a tweet author broadcasts from the affected region at the time of the event (Morstatter *et al.*, 2014; Poblete *et al.*, 2011). The content features of tweets is found to be different for different credibility levels (Mitra *et al.*, 2017b; O'Donovan *et al.*, 2012). However, no previous research has investigated how the distribution of information sources differ within and outside the country of event, and how the author location can impact on linguistic features distribution among credibility level.

Our study found that the location of an author influences the distribution of sources that contribute at the time of an event, and this influence also impacts on the credibility distribution over two different locations. As we see in Table II the distribution of many source types differ across locations. This finding can impact on the users' classification models (De Choudhury *et al.*, 2012) where they do not consider the source based on location. For example, during a crisis, it is necessary to define the type of users along with information categories, as in Imran and Castillo (2015).

Table II shows the source distribution locally and remotely, and Figure 1 presents the informativeness status of different sources. These type of results can provide an expectation of what type of sources and informative status that will contribute in time of event from different locations which can be beneficial to many people. For example, the stakeholders like the decision making people always turn to Twitter when they need to take a decision (Olteanu *et al.*, 2015). Also, it is a common practice for journalists to deal with huge amount of data manually in order to make a news story (Dailey and Starbird, 2014; Heravi and McGinnis, 2015), it is important for them to know the information types that Twitter might provide in different events.

We believe that our research presents the importance of the textual features to identify the credibility in social media, and also shows the influence of author's location on the way of writing. So, these findings can inform wide range of systems, for example, in news reporting systems which try to reach to a credible source in time of event like a journalist or eyewitness (Diakopoulos *et al.*, 2012), or the system for fact checking which differentiate between high and low credible content (Popat *et al.*, 2017). While we do not claim that a stand alone system with only textual features can be set out for rumor detection, but they can be used as extra credibility signals.

Most of the current works that attempt to classify credibility in social media like Twitter uses features beyond the content of the tweets, usually use the social based and network-based approaches (Castillo *et al.*, 2013), temporal approaches (Mitra *et al.*, 2017a) or popularity of tweets (Mendoza *et al.*, 2010). As all of them are useful, all of the previous features need sometimes after posting the content to be collected (Zhao *et al.*, 2015). Using the linguistic features as a marker is a key factor for identifying incredible content, preventing a possible damage that can occur (Bovet and Makse, 2018).

Additionally, the current credibility prediction models face difficulty when they are applied to different events (Boididou *et al.*, 2014; Aker *et al.*, 2017), as the performance accuracy is overestimated. Our results found that there are significant differences of content of the same credibility level and topic when generated from different locations. For example, Table VIII shows many differences of the linguistics features of the same topic but with different location. This finding shows the impact of location on features distribution, which might indicates some ambiguity behind the credibility classifiers variance across topics.

When we look at features proportions of the tweet, we can see some commonalities regardless of credibility. For example, the URL features highly in a remote location in credible and incredible level. This suggests that remote users always share third party information as an external source in their tweets: using URL as an indicator of credibility at the time of an event might not be very accurate because that might relate to the place of the author more than to credibility. However (Popat *et al.*, 2016) shows the impact of the web source reliability hosting such article on increasing the performance of the credibility model prediction, the same approach can be used in the URLs within social media.

The impact of location on the used language for the tweets of the same credibility level. For example, Table IX shows that incredible tweets don't always have the same features when generated from different locations. This is a bit different to (Kwon *et al.*, 2013), as they found that rumors always share the same linguistic features in different topics. On the other side, there are many significant differences between the two location in credible level in topic like crisis, while there was less impact of location on incredible tweets for the same topic. This is an interesting finding, which mean the credible tweets have different characteristics when generated close from the affected region. This is might have an impact on the eyewitness identification research (Morstatter *et al.*, 2014).

As we found that incredible information shares the same linguistic features in social media, they share the same behavior in different contexts like financial fraud and web pages. Humpherys *et al.* (2011) investigated the text of hundreds of financial disclosures and found that fraudulent disclosures use some linguistic features more than non-fraudulent ones such as number of words. The researchers were able to achieve high classification accuracy using linguistic features. On the other hand, in judging website credibility, Wawer *et al.* (2014) used linguistic features to predict the credibility of web pages. They found that trusted words are associated with government web pages for example. The researchers' classification models achieved accurate results.

Findings have shown how textual features have the power to assess credibility, comparing to other features like visual ones, which can lead sometimes to wrong judgments by users (Zubiaga and Ji, 2014). So, implementing a credibility model which can handle the text of different contexts will be beneficial.

In this research we mainly focus on the source and credibility of the content, regardless of information type. The information may be credible but not highly informative, so future research could consider the content informative as (Kumar *et al.*, 2013) and credibility, and study the credibility with time as in some cases like crisis time is very important factor. If we know that a very few users generate most of the social media content, for example, 2 percent of Twitter users generate fifty percent of the tweets (Baeza-Yates and Sáez-Trumper, 2015). That adds further more challenge to find informative and credible source at the same time because of scarcity of rare in the information sources.

Conclusion

In this work for a diverse set of events, we have particularly consider the impact of location on information source and credibility level in social media. By developing a hypothesis driven by previous research that location of users affect their behavior, we included the

location of both event and author in our study to clarify their impact on credibility status. The research questions that we investigated in this paper were:

RQ1. What are the types of sources expected in different events from both in- and outside the country of events?

RQ2. How linguistic features differ among sources of different type, credibility level, and location?

We found that distribution of some sources differ between locations significantly. For the second research question, we found the tweets of the same credibility level have different linguistic features based on their distance from an event and the topic of an event. In future we will include other features in way to have a complete list of common and different behaviors between sources across locations. Moreover, information type with users location can influence content credibility and this is part of our proposed future research.

Notes

1. www.geonames.org/
2. www.merriam-webster.com/dictionary/credible. This definition was used by most of the previous credibility research such as (Castillo *et al.*, 2011). So, applying this definition on our research, a given tweet is said to include credible information, when annotator believe the truthfulness of tweet's information.
3. https://dev.twitter.com/streaming/overview
4. www.crowdflower.com/
5. http://liwc.wpengine.com

References

Aker, A., Zubiaga, A., Bontcheva, K., Kolliakou, A., Procter, R. and Liakata, M. (2017), "Stance classification in out-of domain rumours: a case study around mental health disorders", *International Conference on Social Informatics, Springer*, pp. 53-64.

Baeza-Yates, R.A. and Sáez-Trumper, D. (2015), "Wisdom of the crowd or wisdom of a few?: An analysis of users' content generation", *Proceedings of the 26th ACM Conference on Hypertext & Social Media, HT 2015, Guzelyurt, TRNC, September 1-4*, pp. 69-74.

Bastian, M., Hayes, M., Vaughan, W., Shah, S., Skomoroch, P., Kim, H., Uryasev, S. and Lloyd, C. (2014), "Linkedin skills: large-scale topic extraction and inference", *Proceedings of the 8th ACM Conference on Recommender Systems, ACM*, pp. 1-8.

Bhattacharya, P., Ghosh, S., Kulshrestha, J., Mondal, M., Zafar, M.B., Ganguly, N. and Gummadi, K.P. (2014), "Deep Twitter diving: exploring topical groups in microblogs at scale", *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM*, pp. 197-210.

Boididou, C., Papadopoulos, S., Kompatsiaris, Y., Schifferes, S. and Newman, N. (2014), "Challenges of computational verification in social multimedia", *Proceedings of the 23rd International Conference on World Wide Web, ACM*, pp. 743-748.

Bovet, A. and Makse, H.A. (2018), "Influence of fake news in Twitter during the 2016 US presidential election", CoRR, available at: http://arxiv.org/abs/1803.08491 (accessed May 5, 2018).

Canini, K.R., Suh, O. and Pirolli, P. (2011), "Finding credible information sources in social networks based on content and social structure", *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on Social Computing (SocialCom), PASSAT/SocialCom 2011, Boston, MA, October 9-11*, pp. 1-8, available at: https://doi.org/10.1109/PASSAT/SocialCom.2011.91

Carletta, J. (1996), "Assessing agreement on classification tasks: the kappa statistic", *Computational Linguistics*, Vol. 22 No. 2, pp. 249-254.

Castillo, C., Mendoza, M. and Poblete, B. (2011), "Information credibility on Twitter", *Proceedings of the 20th International Conference on World Wide Web, ACM*, pp. 675-684.

Castillo, C., Mendoza, M. and Poblete, B. (2013), "Predicting information credibility in time-sensitive social media", *Internet Research*, Vol. 23 No. 5, pp. 560-588.

Cheng, Z., Caverlee, J. and Lee, K. (2010), "You are where you tweet: a content-based approach to geo-locating Twitter users", *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM*, pp. 759-768.

Cheng, Z., Caverlee, J., Barthwal, H. and Bachani, V. (2014), "Who is the barbecue king of Texas?: A geo-spatial approach to finding local experts on Twitter", *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'14*, *Gold Coast*, *July 6-11*, pp. 335-344.

Dailey, D. and Starbird, K. (2014), "Journalists as crowdsourcerers: responding to crisis by reporting with a crowd", *Computer Supported Cooperative Work*, Vol. 23 Nos 4/6, pp. 445-481.

De Choudhury, M., Diakopoulos, N. and Naaman, M. (2012), "Unfolding the event landscape on Twitter: classification and exploration of user categories", *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, ACM*, pp. 241-244.

Diakopoulos, N., De Choudhury, M. and Naaman, M. (2012), "Finding and assessing social media information sources in the context of journalism", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM*, pp. 2451-2460.

Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N. and Gummadi, K. (2012), "Cognos: crowdsourcing search for topic experts in microblogs", *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM*, pp. 575-590.

Gibbons, J.D. and Chakraborti, S. (2011), "Nonparametric statistical inference", *International Encyclopedia of Statistical Science*, Springer, Berlin and Heidelberg, pp. 977-979.

Golbeck, J. (2016), "Detecting coping style from Twitter", *International Conference on Social Informatics, Springer*, pp. 454-467.

Gottfried, B.Y.J. and Shearer, E. (2016), *News Use Across Social Media Platforms 2016*, Pew Research Center, Washington, DC.

Gupta, A. and Kumaraguru, P. (2012), "Credibility ranking of tweets during high impact events", *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media, ACM*, p. 2.

Gupta, A., Kumaraguru, P., Castillo, C. and Meier, P. (2014), "Tweetcred: a real-time web-based system for assessing credibility of content on Twitter", *Proceedings, 6th International Conference on Social Informatics (SocInfo)*, *Barcelona, November 11-13*.

Gupta, A., Lamba, H., Kumaraguru, P. and Joshi, A. (2013), "Faking sandy: characterizing and identifying fake images on Twitter during hurricane sandy", *Proceedings of the 22nd International Conference on World Wide Web, ACM*, pp. 729-736.

Han, B., Cook, P. and Baldwin, T. (2014), "Text-based Twitter user geolocation prediction", *Journal of Artificial Intelligence Research*, Vol. 49 No. 1, pp. 451-500.

Hecht, B., Hong, L., Suh, B. and Chi, E.H. (2011), "Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM*, pp. 237-246.

Heravi, B.R. and McGinnis, J. (2015), "Introducing social semantic journalism", *The Journal of Media Innovations*, Vol. 2 No. 1, pp. 131-140.

Horne, B.D. and Adali, S. (2017), "This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news", *The 2nd International Workshop on News and Public Opinion at ICWSM*, pp. 591-600.

Horne, B.D., Nevo, D., Freitas, J., Ji, H. and Adali, S. (2016), "Expertise in social networks: how do experts differ from other users?", *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, May 17-20*, pp. 583-586, available at: www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13125

Humpherys, S.L., Moffitt, K.C., Burns, M.B., Burgoon, J.K. and Felix, W.F. (2011), "Identification of fraudulent financial statements using linguistic credibility analysis", *Decision Support Systems*, Vol. 50 No. 3, pp. 585-594, available at: https://doi.org/10.1016/j.dss.2010.08.009

Imran, M. and Castillo, C. (2015), "Towards a data-driven approach to identify crisis-related topics in social media streams", *Proceedings of the 24th International Conference on World Wide Web, ACM*, pp. 1205-1210.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F. and Meier, P. (2013), "Extracting information nuggets from disaster-related messages in social media", *ISCRAM*, Baden-Baden, May 12-15.

Kang, B., O'Donovan, J. and Höllerer, T. (2012), "Modeling topic specific credibility on Twitter", *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces, ACM*, pp. 179-188.

Kumar, S., Hu, X. and Liu, H. (2014), "A behavior analytics approach to identifying tweets from crisis regions", *Proceedings of the 25th ACM Conference on Hypertext and Social Media, ACM*, pp. 255-260.

Kumar, S., Morstatter, F., Zafarani, R. and Liu, H. (2013), "Whom should i follow?: identifying relevant users during crises", *Proceedings of the 24th ACM Conference on Hypertext and Social Media, ACM*, pp. 139-147.

Kwak, H., Lee, C., Park, H. and Moon, S. (2010), "What is Twitter, a social network or a news media?", *Proceedings of the 19th International Conference on World Wide Web, ACM*, pp. 591-600.

Kwon, S., Cha, M., Jung, K., Chen, W. and Wang, Y. (2013), "Prominent features of rumor propagation in online social media", *13th International Conference on Data Mining (ICDM), IEEE*, pp. 1103-1108.

Magdy, W., Darwish, K., Abokhodair, N., Rahimi, A. and Baldwin, T. (2016), "# isisisnotislam or# deportallmuslims?: Predicting unspoken views", *Proceedings of the 8th ACM Conference on Web Science, ACM*, pp. 95-106.

Mendoza, M., Poblete, B. and Castillo, C. (2010), "Twitter under crisis: can we trust what we rt?", *Proceedings of the First Workshop on Social Media Analytics, ACM*, pp. 71-79.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P. and Rosenquist, J.N. (2011), "Understanding the demographics of Twitter users", *ICWSM*, Barcelona, Catalonia, July 17-21.

Mitra, T., Wright, G.P. and Gilbert, E. (2017a), "Credibility and the dynamics of collective attention", PACMHCI, 1(CSCW): 80:1-80:17, available at: http://doi.acm.org/10.1145/3134715 (accessed October 10, 2017).

Mitra, T., Wright, G.P. and Gilbert, E. (2017b), "A parsimonious language model of social media credibility across disparate events", *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, ACM*, pp. 126-145.

Morstatter, F., Pfeffer, J., Liu, H. and Carley, K.M. (2013), "Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose", *ICWSM*, Boston, MA, July 8-11.

Morstatter, F., Lubold, N., Pon-Barry, H., Pfeffer, J. and Liu, H. (2014), "Finding eyewitness tweets during crises", *Proceedings of the Workshop on Language Technologies and Computational Social Science, ACL, Baltimore, MD, June 26*, pp. 23-27, doi: 10.3115/v1/W14-2509.

Nguyen, D., Gravel, R., Trieschnigg, D. and Meder, T. (2013), " 'How old do you think I am?' a study of language and age in Twitter", *Proceedings of the Seventh International Conference on Weblogs and Social Media (ICWSM), Cambridge, MA, July 8-11*.

O'Donovan, J., Kang, B., Meyer, G., Höllerer, T. and Adali, S. (2012), "Credibility in context: an analysis of feature distributions in Twitter", *2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Conference on Social Computing, SocialCom 2012, Amsterdam, September 3-5*, pp. 293-301.

Oh, O., Agrawal, M. and Rao, H.R. (2011), "Information control and terrorism: tracking the Mumbai terrorist attack through Twitter", *Information Systems Frontiers*, Vol. 13 No. 1, pp. 33-43.

Olteanu, A., Vieweg, S. and Castillo, C. (2015), "What to expect when the unexpected happens: social media communications across crises", *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM*, pp. 994-1009.

Poblete, B., Garcia, R., Mendoza, M. and Jaimes, A. (2011), "Do all birds tweet the same?: Characterizing Twitter around the world", *Proceedings of the 20th ACM CIKM International Conference on Information and Knowledge Management, ACM*, pp. 1025-1030.

Popat, K., Mukherjee, S., Strötgen, J. and Weikum, G. (2016), "Credibility assessment of textual claims on the web", *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, October 24-28*, pp. 2173-2178, available at: http://doi.acm.org/10.1145/2983323.2983661

Popat, K., Mukherjee, S., Strötgen, J. and Weikum, G. (2017), "Where the truth lies: explaining the credibility of emerging claims on the web and social media", *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, April 3-7*, pp. 1003-1012, available at: http://doi.acm.org/10.1145/3041021.3055133

Qazvinian, V., Rosengren, E., Radev, D.R. and Mei, Q. (2011), "Rumor has it: identifying misinformation in microblogs", *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, pp. 1589-1599.

Rosso, P. and Cagnina, L.C. (2017), "Deception detection and opinion spam", in Cambria, E., Das, D., Bandyopadhyay, S. and Feraco, A. (Eds), *A Practical Guide to Sentiment Analysis*, Springer, Cham, pp. 155-171.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2010), "Earthquake shakes Twitter users: real-time event detection by social sensors", *Proceedings of the 19th International Conference on World Wide Web, ACM*, pp. 851-860.

Shariff, S.M., Zhang, X. and Sanderson, M. (2014), "User perception of information credibility of news on Twitter", *Advances in Information Retrieval – 36th European Conference on IR Research, ECIR 2014, Proceedings, Amsterdam, April 13-16*, pp. 513-518, available at: https://doi.org/10. 1007/978-3-319-06028-6_50

Starbird, K. and Palen, L. (2010), *Pass it on?: Retweeting in Mass Emergency*, International Community on Information Systems for Crisis Response and Management, Seattle, Washington, DC, May 2-5.

Starbird, K., Palen, L., Hughes, A.L. and Vieweg, S. (2010), "Chatter on the red: what hazards threat reveals about the social life of microblogged information", *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, ACM*, pp. 241-250.

Tausczik, Y.R. and Pennebaker, J.W. (2010), "The psychological meaning of words: Liwc and computerized text analysis methods", *Journal of Language and Social Psychology*, Vol. 29 No. 1, pp. 24-54.

Thomson, R., Ito, N., Suda, H., Lin, F., Liu, Y., Hayasaka, R., Isochi, R. and Wang, Z. (2012), "Trusting tweets: the Fukushima disaster and information source credibility on Twitter", *Proceedings of the 9th International ISCRAM Conference*, pp. 1-10.

Truelove, M., Vasardani, M. and Winter, S. (2014), "Testing a model of witness accounts in social media", *Proceedings of the 8th Workshop on Geographic Information Retrieval, ACM*, p. 10.

Vieweg, S., Hughes, A.L., Starbird, K. and Palen, L. (2010), "Microblogging during two natural hazards events: what Twitter may contribute to situational awareness", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM*, pp. 1079-1088.

Vosoughi, S., Roy, D. and Aral, S. (2018), "The spread of true and false news online", *Science*, Vol. 359 No. 6380, pp. 1146-1151.

Wagner, C., Liao, V., Pirolli, P., Nelson, L. and Strohmaier, M. (2012), "It's not in their tweets: modeling topical expertise of Twitter users", *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom), IEEE*, pp. 91-100.

Wawer, A., Nielek, R. and Wierzbicki, A. (2014), "Predicting webpage credibility using linguistic features", *23rd International World Wide Web Conference, WWW'14, Seoul, Companion Volume, April 7-11*, pp. 1135-1140, available at: http://doi.acm.org/10.1145/2567948.2579000

Zeng, L., Starbird, K. and Spiro, E.S. (2016), "# unconfirmed: classifying rumor stance in crisis-related social media messages", *Tenth International AAAI Conference on Web and Social Media, Cologne, May 17-20*.

Zhao, Z., Resnick, P. and Mei, Q. (2015), "Enquiring minds: early detection of rumors in social media from enquiry posts", *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, May 18-22*, pp. 1395-1405, available at: http://doi.acm.org/10.1145/2736277.2741637

Zubiaga, A. and Ji, H. (2014), "Tweet, but verify: epistemic study of information verification on Twitter", *Social Network Analysis and Mining*, Vol. 4 No. 1, p. 163, available at: https://doi.org/10.1007/s13278-014-0163-y

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M. and Procter, R. (2017), "Detection and resolution of rumours in social media: a survey", CoRR, available at: http://arxiv.org/abs/1704.00656

Appendix

Description of the task

Informativeness

Please read the following tweet, and check the link inside it if needed, and select the most appreciate category:

(1) The tweet is related to the event and informative, if it includes information about the event that useful and helps you understand what happened:

- @CNN: The Oscars werent afraid to get political https://t.co/pE5zvwe6me; https://t.co/XK2FdGvTF8

(2) The tweet is related to event, but it is not informative, if it mentions to the event but it is not help you understand what happened:

- Did Antoine Fuqua just direct a Walmart short film or am I crazy? Oscars2017.

(3) The tweet is not related to the event:

- Thanks so much i love you Oscars2017 ??

(4) The tweet is not applicable, has some problems like the tweet is not readable, or other issues.

Source

Please read the tweet posted at the time of the Cyclone Debbie 2017 in Australia, check the link inside the tweet if needed, and select the most appreciate source of information as:

(1) A government: information published by the official, such as, police, hospitals, etc.

- @BOM_Qld: Radar loop from the #Mackay radar shows the eyewall and eye of #CycloneDebbie as it tracks toward the coast.

(2) A non-profit organization: information published by administration of non- governmental and not for profit organizations such as Red Cross, UNICEF, etc.

- @RACQOfficial: Don't risk your safety, stay off the roads. #FloodedForgetIt #BigWet #bnetraffic https://t.co/jccsBrBrpb

(3) A business: information published by profit-related business and enterprise:

- @AEMO_Media: We are working with @PowerlinkQLD to prepare for TC Debbie and keeping a close eye on the situation. #CycloneDebbie

(4) Traditional and/or internet media: information that published by sources news organizations, web blogs, such as TV, radio:

- @ABCemergency: All #Brisbane schools to close today as former #CycloneDebbie heads south #bigwet

(5) Eyewitness: information reported by an eyewitness to the event, or from his family, friends, etc.

- So our fence has come down. Tried to save it but it's too windy too strong. https://t.co/fp69EZB2QK #Mackay #bowen #CycloneDebbie

(6) A journalist: associated with an organization or a freelance journalist:

- #CycloneDebbie blew the feathers off a cockatoo. https://t.co/hmELUDw65S

(7) Academic, researcher or specialist: an individual who is working in a university or think-tank:

- @climatrisk: Stay safe FNQ a real frightnener. Hope everything is battened down #CycloneDebbie https://t.co/Ea9z5ASZMl

(8) Politician: an individual who is working in government:

- @AnnastaciaMP: #CycloneDebbie is now crossing the coast between #Bowen and #AirlieBeach. Stay safe everyone. https://t.co/9u2mY2zguY

(9) Digerati: an individual who is popular in area of social media and technology:

- Absolutely APPALLED at @Avis car hire charging us $158 to extend our car rental as #CycloneDebbie is stopping us from reach

(10) Celebrity: individual who is famous for any reason, singer, actor, media presenter, etc (not in technology):

- The team at SDBHQ, and the entire #BlakeArmy is thinking of everyone in North Qld. Stay safe Queenslanders, SDBHQ xo

(11) Ordinary individual: users on Twitter posting updates on their daily life, or non-identified sources:

- Good luck North Queensland. Batten down the hatches. Don't drive in flood waters. Look after each other! STAY SAFE! #cyclonedebbie

All the included sources in this studied like organizations (non-profit, business, etc.) have included their profile locations. The same source type can be local and remote of such event at the same time. For example, "the Red Cross" was engaged from both locations in "Cyclone Debbie Australia":

- Australian Red Cross (local) (@RedCrossAU: Affected by #CycloneDebbie? Let your family know you are ok. Register at https://t.co/NruW5WjXtO https://t.co/EGChoXeh9X).
- Papua New Guinea Red Cross (remote) (@PNGRedCross: High tide in 5 hrs. #RedCross running evac centers. The latest from #TVNZ KimberleeDowns #CycloneDebbie #TCDebbie https://t.co/BvLXnOL7kN).

The same thing for other organizations like government, a government account sometime engage in an international event, for example:

- This tweet about 'Rio2016" from "Road authority" of Uganda government located at (Kampala Uganda): (@UNRA_UG: Wishing #TeamUganda at the #OlympicGames all the entire best! We are your fans and you have our support! Cheers #UGA #RIO2016).
- This tweet about "Cyclone Debbie Australia" from "Met Office Storms" of Uk government located at (Exeter, UK): @metofficestorms: Rainfall radar image of #CycloneDebbie which is slow moving off the coast of #Queensland. Peak wind gust 117 mph at Hamilton Island.
- This tweet about "Italy earth quack" from "Italy UN" of Italy government located at (New York): @ItalyUN_NY: At least 241 dead & 2.5k displaced so far after #ItalyEarthquake. We are very grateful for the solidarity in #NYC

- This tweet about "Rio2016" from "Dept of Sport" of Indian ministry of youth located at (India): @IndiaSports: Indian players/teams event schedule, fixtures for #RioOlympics on Day 4. #Rio2016 #Olympics https://t.co/AvwhX9haVe

The same apply for traditional media, all the included traditional media sources have included their profile's location, for example:

- This tweet from "BBC" at "Rio2016" and their profile location is (London, UK): @BBCWorld: If Michael Phelps was a country https://t.co/wTxvBP8CL5 #Rio2016.
- This tweet authored by "The New Your Times" about "oscar2017" and their profile location (New York City): @nytimes: The #Oscars audience wonders: Who is Gary from Chicago? https://t.co/HDbLOGzn6z.

Credibility

Please read the tweet posted at the time of "the event name," check the link inside the tweet if needed, and determine the credibility level of the tweet as:

- the tweets is definitely credible;
- the tweet seems incredible;
- the tweet is definitely incredible; and
- I cannot decide.

Corresponding author
Suliman Aladhadh can be contacted at: suliman.aladhadh@rmit.edu.au

Event news detection and citizens community structure for disaster management in social networks

Radhia Toujani
ISG Tunis, University of Tunisia, Tunis, Tunisia, and
Jalel Akaichi
College of Computer Science, University of Bisha, Bisha, Saudi Arabia

Received 15 March 2018
Revised 4 July 2018
25 September 2018
Accepted 26 September 2018

Abstract

Purpose – Nowadays, the event detection is so important in gathering news from social media. Indeed, it is widely employed by journalists to generate early alerts of reported stories. In order to incorporate available data on social media into a news story, journalists must manually process, compile and verify the news content within a very short time span. Despite its utility and importance, this process is time-consuming and labor-intensive for media organizations. Because of the afore-mentioned reason and as social media provides an essential source of data used as a support for professional journalists, the purpose of this paper is to propose the citizen clustering technique which allows the community of journalists and media professionals to document news during crises.
Design/methodology/approach – The authors develop, in this study, an approach for natural hazard events news detection and danger citizen' groups clustering based on three major steps. In the first stage, the authors present a pipeline of several natural language processing tasks: event trigger detection, applied to recuperate potential event triggers; named entity recognition, used for the detection and recognition of event participants related to the extracted event triggers; and, ultimately, a dependency analysis between all the extracted data. Analyzing the ambiguity and the vagueness of similarity of news plays a key role in event detection. This issue was ignored in traditional event detection techniques. To this end, in the second step of our approach, the authors apply fuzzy sets techniques on these extracted events to enhance the clustering quality and remove the vagueness of the extracted information. Then, the defined degree of citizens' danger is injected as input to the introduced citizens clustering method in order to detect citizens' communities with close disaster degrees.
Findings – Empirical results indicate that homogeneous and compact citizen' clusters can be detected using the suggested event detection method. It can also be observed that event news can be analyzed efficiently using the fuzzy theory. In addition, the proposed visualization process plays a crucial role in data journalism, as it is used to analyze event news, as well as in the final presentation of detected danger citizens' clusters.
Originality/value – The introduced citizens clustering method is profitable for journalists and editors to better judge the veracity of social media content, navigate the overwhelming, identify eyewitnesses and contextualize the event. The empirical analysis results illustrate the efficiency of the developed method for both real and artificial networks.

Keywords Hierarchical clustering, Risk assessment, Social network analysis, Event detection, Citizens' community structure, Social news

Paper type Research paper

1. Introduction

The use of social media has dramatically increased in recent years, which produced novel networked publics for citizen-generated content considered as an essential source of data used as a support for professional journalists (Deborah *et al*., 2017). In fact, the increase of employing social media professionals has become important to document news during crises, war zones, political elections, sports events, etc. (Brandtzaeg *et al*., 2016; Stephens-Davidowitz and Pinker, 2017). Most existing works mining social media for event detection and news gathering addressed the problem of noise and burst detection. Indeed, detecting and filtering non-relevant or noisy content is important for isolated users to generate timely and relevant content. Burst detection issue plays an important role in complex analysis. Moreover, community analysis

This paper forms part of a special section "Social media mining for journalism".

Online Information Review
Vol. 43 No. 1, 2019
pp. 113-132

1468-4527
DOI 10.1108/OIR-03-2018-0091

technique is generally used to enhance events filtering. Visualization is also essential in data journalism as the enormity of information needs to be systematically organized into a format that is easy to communicate. In this context, we develop, in this study, a practical tool to allow journalists and news readers to recognize newsworthy topics based on message streams without being plagued by carrying out a dependency analysis between all extracted news. In fact, most of the proposed research works dealt essentially with the predefinition of a set of keywords and did not investigate the integrity of these terms, which made difficult, for journalists, to ensure the accuracy of the reports gathered from social media. Another complication is that declaring and depicting event news, which may be fuzzy, unclear, incorrect and incomplete, by citizens' social network can produce various difficulties concerning the extracted news quality and meaning. To solve the limits associated with verifying event content and its sources, we integrated the fuzzy theory into a natural language processing (NLP) method to detect both known and unknown news events, check information and its sources and contextualize the event. Besides, the majority of the existing techniques focus on identifying and classifying news events manually, which degrades considerably the effectiveness of event detection process. To deal with the previously-mentioned issues, our technique relies on the burst and noise event detection to isolate non-relevant users with a low degree of danger and to merge citizens generating timely and relevant content. The proposed clustering event method allows journalists to better judge the event news content. The performed studies introduced tools having dashboard-style interfaces with complex data graphics, which is so attractive for some professional users. Thus, visualization process is suggested, in this paper, to detect event news using of power BI dashboards.

The remainder of this paper is organized as follows: In Section 2, we present the literature review about using social media to gather news by expert journalists. Then, in Section 3, we illustrate our employed approach to evaluate both natural dangerous event news detection and citizens' clustering method. Finally, the experiments are demonstrated and a brief description of the visualization of the obtained results is provided.

2. Related areas

The importance of the event detection task in gathering news from social media (Dou *et al.*, 2012) was clearly shown in the literature. Sayyadi *et al.* (2009) designed a community model in order to discover events using keywords, noun, phrases and named entities. Phuvipadawat and Murata (2010) demonstrated the key value of text messages by counting retweets and detecting popular terms like nouns and verbs. Their study was further extended by applying a simple tf-idf scheme employed to specify concepts similarity (Phuvipadawat and Murata, 2011). Afterwards, entities were identified by utilizing the Stanford Named Entity Recognizer to determine communities and similar message groups. Twitcident was introduced, in Abel *et al.* (2012), to broadcast emergency services, identify incidents, extract a set of related keywords and finally gather related updates from social media. Ozdikis *et al.* (2012) detected events employing hashtags through clustering them and identifying semantic similarities among hashtags. Shi *et al.* (2017) depicted a new cosine measure-based event similarity detection technique to evaluate the correlation between events.

Moreover, generated content displayed on social media is so significant in the procedure of capturing news events in order to classify and verify stories. In this context, we can mention the work in Boididou *et al.* (2018) where authors compared three verification methods. The first technique focused on employing textual patterns for the extraction of claims about whether a tweet is fake or real. The second method is based on using information in which the credibility proves the similarity of event news topic extracted from tweets. The third technique applied semi-supervised learning scheme that affects the decisions of two distinct event news credibility classifiers. Experiments performed on

varieties of datasets showed that the third approach outperformed the two other techniques. Besides, content news verification issue was considered as rumors spreading in social media (Derczynski *et al.*, 2017; Enayet and El-Beltagy, 2017).

More recently, filtering non-relevant or noisy content has attracted the attention of the research community. In fact, noisy and non-relevant groups refer to the overwhelming and the inadequacy of group content. Therefore, noise detection is the manual process of checking detecting and filtering non-relevant (or noisy) content. Relevance could be applied to produce either completely automatic approach with no user involvement or semi-automatic methods where the limited involvement from the end user is solicited. In the introduced citizens' clustering method, relevance of detected groups is proved by analyzing event news content through the introduced top-down hierarchical clustering process. In Wei *et al.* (2017), an approach was introduced for filtering non-relevant users according the estimation of the location of Twitter users.

Another problem of news gathering is burst detection which consists in discovering bursts or unexpected rises in frequency of topic and/or locating certain micro-blog events for the purpose of identifying events in case where a given event or news story is trending. TopicSketch was applied, in Wei *et al.* (2016), to detect burst topics in a real-time fashion from a large-scale data. Researchers developed, in Khan *et al.* (2013), an alternative graph-based approach to detect event news. They first applied LDA for the identification of discussion topics within a stream. Subsequently, a graph with the tweets in each topic was built relying on word co-occurrences. Ultimately, PageRank was utilized to identify, in each topic, the salient tweets. In more recent work (Meladianos *et al.*, 2018), another graph-based approach has been introduced to analyze rapid changes occurrence within the graphs as a proxy for the purpose of identifying important events to be incorporated in the summary.

Many works were performed to classify disaster-related tweets. For instance, in Stowe *et al.* (2016), researchers developed an annotation schema containing seven types of tweets and employed a Support Vector Machine (SVM) classifier to identify natural risks from Tweeter. Verma *et al.* (2011) constructed classifiers to identify tweets demonstrating situational awareness in four datasets (Red River floods of 2009 and 2010, the Haiti earthquake of 2010 and Oklahoma fires of 2009). Garg and Kumar (2016) provided an overview about the used approaches analyzing attractive event from social media data including natural disasters data, weather data, temporal data, geo-location data, etc. Garg and Kumar (2016) proved that delivering the exact news event in the appropriate time may improve the existing emergency management systems and minimize risks by providing quick relief intervention.

Besides, many attempts were made to put together dashboards, especially those designed for news gathering from social media, as it is the case of TwitInfo (Marcus *et al.*, 2011) which was used to permit journalists to track the latest news of events in a real time. In Diakopoulos *et al.* (2012), a dashboard for exploration of users was employed for the identification of the various types of users reporting about a specific event. Zubiaga *et al.* (2013) introduced a dashboard for detailed use of news reports taken from Twitter's trending topics. Researchers (Lin *et al.*, 2016) proposed an analogous application, whose visualizations focused more on the social network and the interactions among users during breaking news events on the content of the tweets. In the above-cited studies, researchers generally took into account either NLP or clustering and classification methods to allow journalists navigating the overwhelming amount of user generated content for the purpose of detecting both known and unknown events news, checking data and their sources, identifying eyewitnesses and contextualizing the event and news coverage. However, the fuzzy theory was not examined in order to check textual content related to news. Basically, the process of event detection involves the transformation of unstructured news losing relevant information in the original format. In this sense, it is important to deal with vagueness and uncertainty of real knowledge. The meaning of textual event news content

varied according to application domains. Besides, event news subjectivity clues used with objective senses constitute a significant source of error event news detection. Thus, by exploiting the advantages of fuzzy approaches, it is possible to develop more powerful mechanisms to represent event news because such mechanisms include the verification of uncertainty concepts in the final event news detection model and the issue of word sense disambiguation. Event news requires uncertain knowledge as any event detection task involving imprecise concept; most of the concepts that we are going to use as degree of citizens' danger are linguistic (close, strong, very strong, low, medium, etc.). Thus, fuzzy logic seems to be a natural choice dealing with linguistic variables.

3. Proposed approach

The proposed citizen classification method is divided into three main steps. It starts by cleaning the news extracted online. The second step aims at defining the degree of disaster using the fuzzy theory. In fact, this step is composed of three sub-steps: identification of fuzzy terms; determination of membership function; and computing citizens' danger degree (CzDD). The third step consists in finding groups of citizens with close degrees of danger (Figure 1).

3.1 Event news extraction process

Natural hazards event news extraction can be divided into several sub-tasks; we can consider our event detection process as a pipeline. In our context, event news is made up of three main components:

(1) event news trigger: this is the term that accurately describes the occurrence of an event;

(2) event news arguments: are the particulars involved in an event (participants), i.e. those having related entities event trigger; and

(3) event news mention: is a phrase or sentence (news) in which an event is described. It includes triggers and arguments.

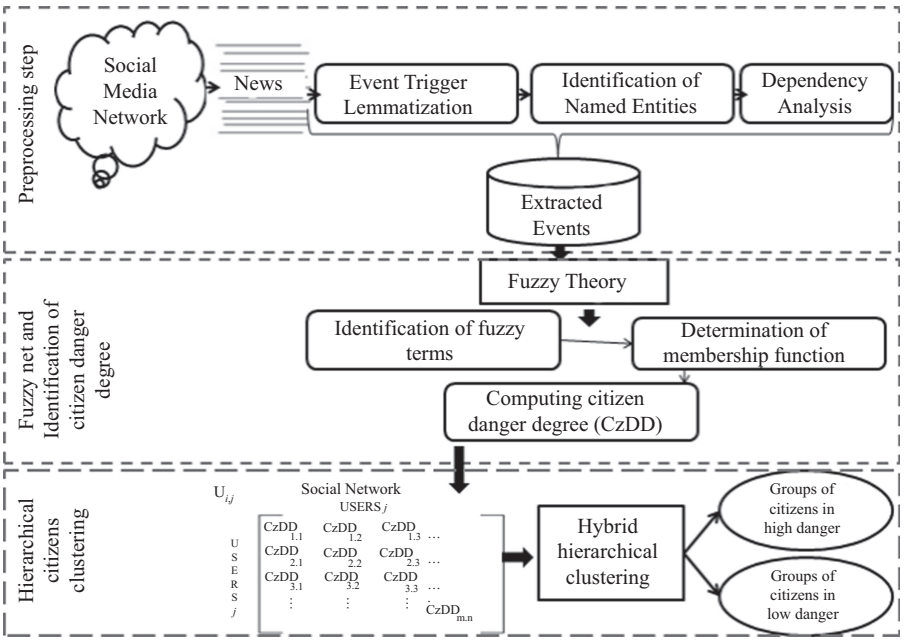


Figure 1.
General proposed modeling framework

The introduced method of extracting events from news starts first by determining the event mention which consists of an event trigger and zero or more arguments. In the second step, we need to extract the different arguments such as the date of the event news and other concerned participants. Consequently, the proposed event extraction process is divided into three main stages:

(1) identification and extraction of events triggers related to natural disasters and based on a created lexicon containing a set of possible natural disasters event trigger;

(2) identification of named entities related to the identified triggers events news; and

(3) dependency analysis between the extracted entities and the event news name (event trigger). We realized, in this study, a semantic-syntactic dependency analysis.

3.1.1 Events triggers detection and recovery. In order to extract events related to natural hazards and contained in a non-structured text, we applied, initially, the lemmatization technique to find the canonical form of words. In fact, we created a list of lemmas considered as possible triggers of potential events describing natural hazards. We used, thereafter, a lemmatizer to annotate our unstructured textual data; we applied, for our environment, the lemmatizer TreeTagger within the TC project[1].

We consider the following example extracted from unstructured social news, "In 2013, an anonymous ignited a forest in the north of Tunisia." TreeTagger applied to this example gave the results presented in Figure 2. In fact, TreeTagger returned the canonical form of each word. For example the word "ignited" (verb) becomes "ignite" (Lemma). Then, based on the annotation results generated by TreeTagger, we browsed the text content. For each time, we detected a lemma appearing simultaneously as well as the list of lemmas (created lexicon) and treated text. Subsequently, we recuperated the word matching each lemma as a possible event trigger related to a possible natural hazard or an event name.

3.1.2 Identification of named entities related to the identified events news triggers. The recognition of named entities is a sub-task of information extraction. We identify usually named entities as all text objects referring to the names of people, places and companies presented in given news. In this current work, we used a customized version of Stanford Named Entity Recognizer in order to enhance the accuracy of named entities recognition (NER).

The application of the customized NER system on the same example provided the following result.

In < DATE > 2013 < /DATE > < PERSON > an anonymous < PERSON > ignited a forest < LOCATION > in the north of Tunisia < /LOCATION > .

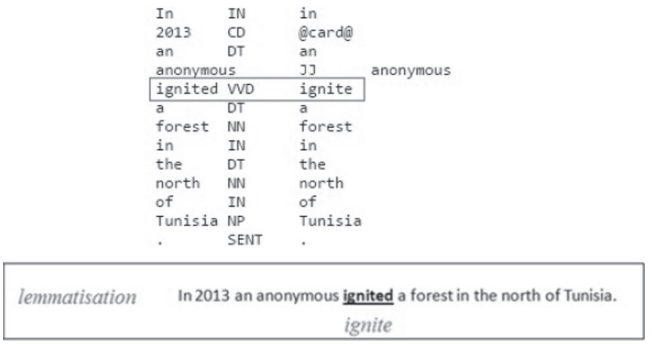


Figure 2.
Results obtaining by using TreeTagger

We notice that the system recognized three different named entities < DATE >, < PERSON > and < LOCATION >, which are considered in the introduced method as the event's participants related to the main event news trigger. However, we still needed to prove that these extracted entities are related to the main event news trigger in order to confirm the coherence of the extracted information. For this reason, we proposed to achieve a semantic and lexical dependency analysis or parsing.

3.1.3 Dependency analysis between the extracted named entities and recuperated event news triggers. After extracting the necessary news, we should show that the extracted data are linked to the main event. In other words, we must prove that there is a dependency relation between the different extracted information and the detected main event trigger. To perform this work, we used a personalized version of Stanford Parser. The system generated a description of the grammatical relations between words in a sentence. Then, we exploited this description to determine if there is any dependency links between the various constituent terms of the analyzed sentence, including extracted entities, and events triggers.

3.2 Part 2: applying fuzzy sets theory on the extracted events news: Computing citizens degree of danger

3.2.1 Step 1: identification of fuzzy concepts. The objective of this step is to clarify the fuzzy parts in the extracted events news process. Indeed, we defined the fuzzy parts of the list of detected events news; these events were stored in a thesaurus (*THS*) containing all the events extracted through the application of the proposed extraction process outlined in the previous section. After a detailed analysis of extracted events, we recovered the different fuzzy concepts into a new dictionary of fuzzy terms named (*DFT*) based on a particular value of a linguistic variable. Thus, the linguistic variables are terms used during the description of a situation, a phenomenon or process, such as temperature, speed, etc. The values of the linguistic variables are linguistic translation of the various states of the latter. For example, low, medium and high are values of the temperature linguistic variable. In fact, the definition of a linguistic variable includes both numerical and linguistic information. The linguistic terms were considered as fuzzy sets of the universe of discourse using numerical values.

After extracting the list (*THS*), we generated a base of the linguistic variable intended to natural hazards. The basis of the variable was used later to identify the fuzzy concepts.

The phase of determining of fuzzy concepts took, as input, all extracted events. In this step, we verified the existence of the concept in the fuzzy terms base which would output the list of vague concepts.

Table I shows samples of fuzzy concepts for the case of two natural hazards: earthquake and storm.

3.2.2 Step 2: computing CzDD: membership functions. We associated membership functions to fuzzy terms extracted from social news. The most used membership functions are: the triangular function, monotonically-increasing function, monotonically-decreasing function and trapezoidal function (Zadeh, 1999). In this work, we used the first threefunctions: monotonically-increasing function, monotonically-decreasing function and triangular function. We did not employ trapezoidal function as it generates membership degrees generally of the value 1. Thus, each term would have its own membership function.

Table I.
Fuzzy concepts of two natural hazards: earthquake and storm

Event type	Fuzzy terms
Earthquake	Micro, very minor, minor, light, moderate, strong, very strong, major, devastator
Storm	Low, moderate, strong, very strong, devastating

The degree of citizens' danger was obtained as follows:

$$CzDD_T(EN) = \begin{cases} 0, & \text{if } EN \leqslant \ell \text{ ou } EN \geqslant \beta \\ \frac{EN-\ell}{\alpha-\ell}, & \text{if } \ell < EM \leqslant \alpha \\ \frac{\beta-\ell}{\beta-\alpha}, & \text{if } \alpha < \ell < \beta \end{cases}, \quad (1)$$

where EN is the extracted event news, ℓ denotes a low value of EN, β represents high value of EN and α is the modal value.

Tables II and III show the partition of fuzzy functions for each fuzzy event for earthquakes' description.

3.2.3 Step 3: calculation of membership degrees. For the proposed method, the purpose of calculating of membership degree between the fuzzy events and fuzzy concepts is to determine citizens' degree of danger for each extracted event news.

Example 1:

"According to the records analyzed yesterday, wind gusts reached 138km/h in Montauban, Monday at 8:12y pm."

138 km/h $\in$ [100, 180]: triangular function $Czdd_{EN}(Wind{-}Montauban) = ((138-100)/(150-100)) = 0.76$

$\Rightarrow$Wind speed in Montauban presents a moderate danger with a degree of 0, 76.

Example 2:

"In the north of the island of Balagne in Cap Corse, in the late afternoon of Saturday, the wind will reach 140–170 km/h. At midnight, it will turn north again with violent gusts across the north of the island."

170 km/h $\in$ [150, 200]: triangular function

$CzDD_{DF}(Wind{-}Balagne) = ((170-150)/(180-150)) = 0.7$

$\Rightarrow$The wind speed in the island of Balagne presents a strong danger with a degree of 0, 7.

Example 3:

9 km/h $\in$ [0, 150]: monotonically-decreasing function

$CzDD_{EN}(Wind{-}Tozeur) = 9 < 150$

$\Rightarrow$The wind speed in Tozeur presents a Low danger with a degree of 1.

Micro	$\leqslant$1, 9	–	–	Monotonically-decreasing function
Very minor	2	2.5	2.9	Triangular function
Minor	3	3.5	3.9	Triangular function
Light	4	4.5	4.9	Triangular function
Moderate	5	5.5	5.9	Triangular function
Strong	6	6.5	6.9	Triangular function
Very strong	7	7.5	7.9	Triangular function
Major	8	8.5	8.9	Triangular function
Devastator	$\geqslant$9, 0	–	–	Monotonically-increasing function

Table II. Fuzzy concepts for natural hazard: earthquake

Low	$\leqslant$100	–	–	Monotonically-decreasing function
Moderate	100	150	180	Triangular function
Strong	180	200	210	Triangular function
Very strong	210	220	250	Triangular function
Devastator	$\geqslant$250	–	–	Monotonically-increasing function

Table III. Fuzzy concepts for natural hazard "storm"

3.3 Step 3: hierarchical citizens' clustering

Community analysis can be used to filter events and to develop a computational actuality model to offer automatically better story. To ensure the veracity assessment, we proposed citizens' community analysis method as indicator of veracity. In fact, we modeled social network by a graph $G = (V, E, D)$, where V represents the citizens in media networks, E denotes the different interactions between them, namely, the sets of exchanged news, and D is the degree of danger obtained by applying the introduced fuzzy method. We weighted every edge by the value of *CzDD* defined in Equation (1).

Then, we proposed a mixed citizens clustering method which combines hierarchical classification operators, namely agglomerative and divisive operators. In fact, agglomerative technique is based on the aggregation process which gathers two citizens having the highest *CzDD* value, while the divisive operator relies on decomposition process based on dividing a group of citizens into two sub-groups having the lowest *CzDD* value.

For the introduced descendant or top-down process, citizens' communities were iteratively split into two parts by removing edges between the least dangerous vertices. However, ascendant or bottom-up process began by the fact that each node or citizen constitutes a separated community and ends up with considering the whole graph as one community.

Therefore, the suggested mixed method requires the existence of an initial partition of k citizen community. It proceeds by a successive combination of the ascendant and descendant operators and ends up if a stable partition is obtained by applying either aggregation or decomposition operator.

3.3.1 Initial community structure. The proposed mixed citizens classification method (*MCCM*) is based on the assumption that an initial solution is originally composed of n partitions (groups). It does not change the number of partitions, but it modifies the initial distribution. In our case, the generation of the initial community structure was obtained using the method suggested in Toujani and Akaichi (2017), which considers the initial partition as a combinatorial optimization issue and solves it by employing Tabu Search metaheuristic. Therefore, the initial generated citizens' partition was injected as input to *MCCM*.

3.3.2 Bottom-up citizens' clustering algorithm (B−upCCA). At the beginning, each vertex represents a separate community. *B−upCCA* initially considers that each citizen (social network user) constitutes a community: $C = \{\{Cz_1\}, \{Cz_2\}, \ldots, \{Cz_n\}\}$. At each iteration, the introduced *B−upCCA* moves from one level n to the next one by performing an aggregation procedure which consists in clustering two citizens, having the highest degree of danger, in the same community. The algorithm ends up if all citizens are clustered in the same community and the aggregation procedure is no longer feasible.

In fact, for a given level n, *B−upCCA* examines all pairs of groups constituting the community C_k to find the two groups c_α and c_β having the highest *CzDD* value. The pseudo-code of *B−upCCA* is given in Algorithm 1.

Algorithm 1. *B−upCCA*
Require: input: weighted graph $G(V, E, D)$
Ensure: k sub-detected citizens groups
1: $K := n-1$
2: **repeat**
3: $C_k = C_k \backslash \{c_\alpha, c_\beta\} \cup \{c_\alpha \cup c_\beta\}$ such that.
$CzDD(c_\alpha, c_\beta) = \max(CzDD_{i,j}(cz_i, cz_j))$.
$CzDD(C_k) := CzDD(C_k - CzDD(c_\alpha, c_\beta))$ $k := k-1$
4: **until** $K = 2$

where C_k represents the k generated citizens groups, c_k denotes the k sub-detected group, cz the social network citizen' constituting c_k and n corresponds to the relative agglomerative hierarchical level.

3.3.3 Top-down citizens clustering algorithm (Top−DCCA). The introduced *Top−DCCA* is based on descendant process and decomposition principle. In fact, the introduced *Top−DCCA* is similar to the *B−upCCA*. But, it is proceeds by an opposite hierarchical construction relying on the successive separations of citizens' community into two groups having the least important *CzDD*. Indeed, *Top−DCCA* initially considers that all social network citizens constitute a community and moves from one descendant hierarchical level to the next one by performing a decomposition procedure: successive separations into two citizens' group having the least *CzDD* value. The stop condition is achieved when the bursting procedure is no longer feasible and each social network citizens constitutes a separate group.

Therefore, in Algorithm 2, we describe the pseudo-code of the proposed *Top−DCCA*:

Algorithm 2 *Top−DCCA*
Require: input: graph *G(V, E, D)*
Ensure: output: k sub-detected citizens communities
1: $K: = 1$
2: **repeat**
3: $C_k = C_k \backslash g_k \cup g_k \backslash cz^*$, cz^* Such that.
$(g_\alpha \backslash m^*, m^*) < Score_{CzDD}(c_\alpha \backslash cz, cz) \forall g_\alpha \in C_k, \forall cz \subset c_\alpha$.
$CzDD(Top-DCCA(C_K)) = CzDD(C_K + CzDD(g_\alpha \backslash cz^*, cz^*))$ $k = k+1$
4: **until** $K = n$-1

where C_k represents the k generated citizens community, g_k denotes the k sub-detected citizens group at each hierarchical level, cz corresponds to the social network citizens constituting g_k and n is the related hierarchical level.

The number of partitions explored by *Top−DCCA* is more important than that explored by *B−upCCA*. In fact, to find the best division of a community with p partitions into two sub-partitions, we need to explore $2^{p-1}-2$ possibilities.

3.3.3.1 Mixed Hierarchical Citizens Clustering method (MHCCM). The introduced *MHCCM* proceeds by a successive combination of *Top−DCCA* and *B−upCCA* techniques. In fact, the divisive process relying on the decomposition operator provides a partition of $(k-1)$ groups on which it is sufficient to apply the aggregation operator in order to obtain the number of the initial group (K) and vice versa. *MHCCM* ends up if the obtained sub-detected communities remain constant. Thus, the process of stabilization is achieved and we get the same detected citizens' community by applying either bottom-up method or top-down technique.

Hence, the pseudo-code of the introduced *MHCCM* is described in Algorithm 3:

Algorithm 3 *MHCCM* (C_k)
Require: input: initial citizens community structure (C_k)
Ensure: output: mixed hierarchical citizens community structure (C'_k)
repeat
repeat
$C_k = B\text{-}upCCA \circ Top\text{-}DCCA(C_k)$.
until $(B-upCCA \circ Top-DCCA(C_k)) = C_k$
repeat
$C_k = Top\text{-}DCCA \circ B\text{-}upCCA(C_k)$.
until $(Top-DCCA \circ B-upCCA\ (C_k)) = C_k$
until $(B-upCCA \circ Top-DCCA(C_k)) = (Top-DCCA \circ B-upCCAC_k)) = C_k$

4. Experimental evaluation

In this section, we validate and evaluate the performance of the proposed method for news event detection and citizen community structure:

- Data set description: the evaluation of our framework was conducted on both real and artificial networks in order to obtain complete and decisive results.
- Real networks: to assess the performance of the introduced news event detection model, we used the CrisisLexT26 data set (Olteanu *et al.*, 2014) including tweets collected during 26 crisis events that happened in 2012 and 2013. Each crisis contains around 1,000 annotated tweets for a total of around 28,000 tweets with labels indicating whether a tweet is related to a crisis event or not. More information about the CrisisLexT26 data set can be found on the CrisisLex website[2].
- Artificial networks: we used the Social News On the Web (SNOW) 2014 Data Challenge which allowed creating a public benchmark and evaluating resource for the problem of topic detection in streams of social content (Papadopoulos *et al.*, 2014). This benchmark was employed to retrieve newsworthy stories or topics, for multiple timeslots over 24 h, where each timeslot lasted 15 min.
- Baseline approaches: the eficiency of the proposed method was validated by comparing our model against Naves Bayes and SVM as baseline approaches. As the primary goal of our work is to capture tweets related to natural hazard event, the performance of the introduced method was compared with the work of Verma *et al.* (2011).
- Performance assessment: the choice of a suitable evaluation measure depends on the cluster properties. Therefore, to show the homogeneous distribution of groups, we used cross-entropy clustering (*CEC*) (Spurek, 2017) and Precision-Recall as evaluation models. Besides, to prove that the obtained citizens' clusters are compact and well separated, we employed Davies–Bouldin Index (*DBI*) (Moshtaghi *et al.*, 2018), which allowed providing clusters with the minimum intra-cluster distance as well as the maximum distance between cluster centroids. The minimum value of the index indicates a suitable partition of the data set. Besides, we used Silhouette index (Martinez *et al.*, 2017) to compute, for each object, a width depending on its membership in any cluster. Moreover, there are two main categories of testing criteria: external indices and internal indices (Rui, 2012). In fact, the former are distinguished by the presence of priori information of known categories. For this reason, we used CEC and Precision-Recall as external clustering quality criteria for SNOW benchmark because the latter has a model partition and a labeled data challenge. However, when the quality measures are based on the examples, we applied silhouette plot and DBI as internal criteria for the CrisisLexT26 data set.

4.1 Artificial networks

Topics extracted from SNOW benchmark were evaluated on several dimensions, namely, precision and recall, as well as coherence, relevance and diversity. In fact, to prove the coherence and relevance of citizen' clusters, we applied an entropy-based criterion (*CEC*) (Spurek, 2017) as clustering quality criterion.

4.1.1 Clustering quality relying on Precision-Recall. In Figure 3, we compare the performance (in terms of Precision/Recall) of SVM and Naves Bayes classifiers and Verma' method to that of the introduced event news detection and citizens' community structure method.

From the above-presented curves, we notice that the development of news event method was able to successfully identify danger citizen' community. We also remark that our approach performed almost perfectly (with Precision > 0.89) and its quality is generally better than that

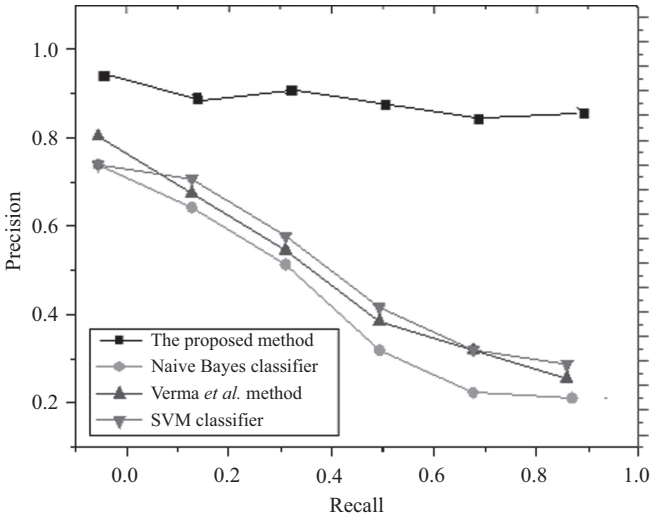


Figure 3.
Comparison of the clustering quality in terms of Precision-Recall

of SVM, Naves Bayes and Verma' classifiers due to the high efficiency of the introduced mixed hierarchical clustering. In fact, because the introduced mixed process allowed adjusting the number of clusters, the hierarchical citizen classification process was more flexible. Naves Bayes classifier had high recall rate with lower precision (with Precision = 0.22 and Recall = 0.89) indicating a worst classification performance. In addition, Verma' method and SVM classifier provided unbalanced danger citizens' classes for identifying crisis-related tweet with precision = 0.35 and Recall = 0.70. Obviously, the performance of baselines approaches significantly dropped when identifying groups of danger citizens'.

Overall, the suggested mixed citizens' clustering method appears as the best model for identifying danger citizens' community and crisis-related tweet. Furthermore, the development of danger citizen classification generated balanced citizens' classes and provided a balance between Precision and Recall with higher precision to higher recall values (> 80).

4.1.2 Clustering quality relying on CEC. In this work, we used *CEC* as entropy-based density criterion to verify the efficiency of the introduced mixed hierarchical citizens' cluster and to find the optimal number of clusters by automatically removing citizens' groups having negative information cost.

Let $C = c_1, c_2, \ldots, c_k$ be a partitioning computed by any model. The elements of each cluster c_k were coded by an optimal density d_i out of a family D_i:

$$CEC(c_1, d_1; \ldots; c_k, d_k) = \sum_{i=1}^{i=k} p_i.\left(-\ln(p_i + H^{\times})(C_i||d_i)\right), \tag{2}$$

where $p_i = (c_i/c)$, and $H^{\times}(C_i||d_i)$ denotes the most common density class for which the cross-entropy is:

$$H^{\times}(y||G\Sigma)) = \frac{N}{2}\ln(2\pi) + \frac{1}{2}tr\left(\Sigma^{-1}\Sigma_x\right) + \frac{1}{2}\ln\det(\Sigma). \tag{3}$$

Generally, *CEC* decomposed m members into k clusters to minimize the cost function called energy E of the clustering by switching the members between clusters. *CEC* function

reduced clusters whose cardinalities decreased below a given small prefixed level. Indeed, the energy function or cost function E is given by:

$$E(c_1, D_1;\ ...;\ c_K, D_K) = \sum_{i=1}^{k} p(c_i)(-\ln(p(c_i))) + H^{\times}\left(c_i \middle|\middle| D_i\right), \tag{4}$$

where c_i stands for the $i-th$ cluster, $p(c_i)$ is the ratio of the number of members in ith cluster to the total number members and $H(c_i||D_i)$ corresponds to the value of cross-entropy representing the internal cluster energy function of data c_i defined with respect to a certain density family D_i encoding the considered type of clustering.

Figure 4 represents a comparison of the performance of baselines methods, in terms of *CEC* values with that of the introduced citizens' clustering technique.

As shown in Figure 4, the cost function of the introduced citizens' clustering method is lower than the energy function of baselines techniques. In fact, the energy function of the proposed method increases at the beginning of iterations. Then, it decreases from iteration 15, indicating better clustering quality. Nevertheless, energy function value of Nave Bayes classifier rises proportionally with the number of iterations increase showing the worst quality of clustering when identifying danger citizens' groups. For the SVM classifier, the energy function drops significantly to detect danger event news (with 1.25 cost function at iteration 20). Since the number of iterations is higher than 20, the energy function value rises, reflecting that SVM classifier is the worst performing compared to the other methods.

It is also noticed that the cost function of Verma' method degrades polynomially with the increase in the number of iterations. But, it is higher than the energy function of other methods, revealing the bad classification performance.

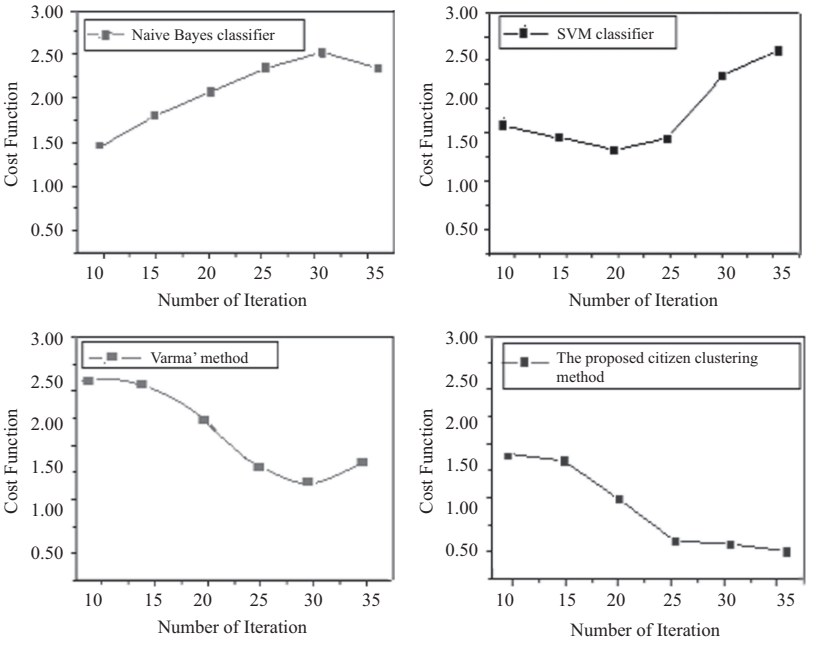


Figure 4.
Comparison of the clustering quality in terms of *CEC* value

Overall, we see that the introduced method displayed a quite low cost function regardless of the number of iterations, indicating better clustering results. Indeed, our citizens clustering technique performed almost perfectly and its quality is better than that of the baselines methods, specifically with the increase of the number of iterations (with cost function $\leqslant 1.0$ and number of iteration varying between 20 and 30 iterations).

Figure 5 depicts the distribution of the proposed mixed citizens clustering method, provided by *CEC* function, compared to that obtained by Nave Bayes, SVM and Verma' classifiers when identifying four danger citizens' clusters.

As demonstrated in Figure 5, the distribution of citizen' clusters is based on the density obtained for each cluster over a timeslot. By comparing the proposed method, with baselines approaches, we notice that the introduced mixed hierarchical citizens clustering method displayed better partitioning and produced denser citizens' communities in a shorter time. Consequently, as seen in histogram presented in Figure 5, *CEC* puts the barrier in a more reasonable place. For example, the third cluster obtained by our technique has the best dense partitions with 0.08 density value.

The difference between the distribution of citizens' tweets per timeslot of Nave Bayes and SVM classifiers is negligible, with the exception of the second cluster where Naves Bayes provided the lowest denser citizen' groups. Besides, Verma' method displayed quite a low density regardless of the time.

4.2 Real network

To show that the obtained citizens' groups are compact and well separated, we measured the quality of the citizens' clustering results, for real network, in terms of silhouette index (Martinez *et al.*, 2017) and *DBI* (Moshtaghi *et al.*, 2018).

4.2.1 Clustering quality relying on silhouette value. Silhouette statistics represents efficient internal performance criteria for clustering. It was calculated for each sub-partition to see the adaptation between a specific sub-partition and the cluster assigned to it.

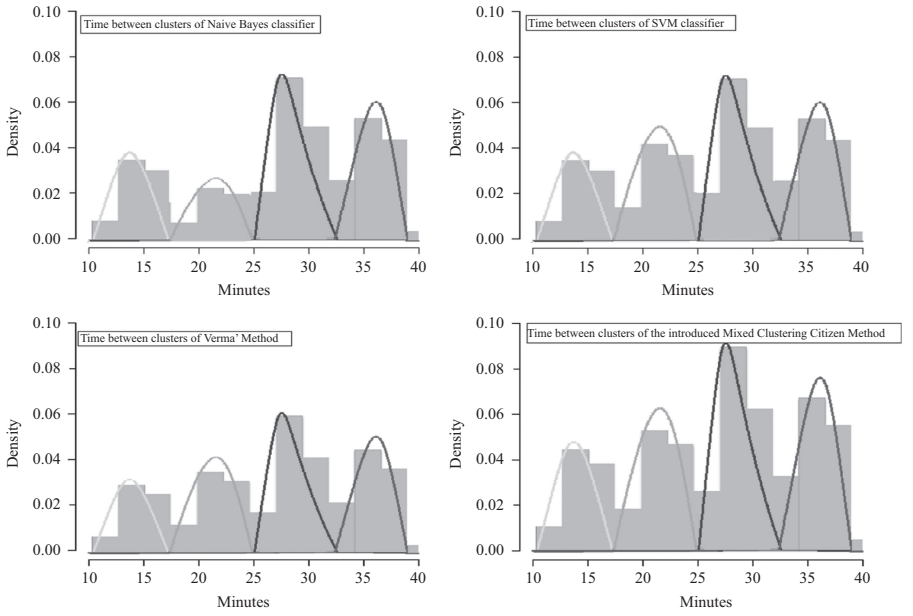


Figure 5.
Histogram with *CEC* density approximation provided by four clusters over a timeslot

Thus, silhouette index was obtained by comparing how close the sub-partitions are to other sub-partitions in its own cluster with how close they are to sub-partitions in other clusters:

$$\text{Let } C = \underbrace{\{\{cz_1, cz_2, cz_3\},\}}_{p1}, \underbrace{\{\{cz_4\}\}}_{p2}, \ldots, \underbrace{\{\{cz_n\}\}}_{pi}.$$

The obtained cluster C is made up of i sub-partition p composed of n citizens cz.

Each p_i is represented by a silhouette displaying citizens which lays well within the partition and those which are marginal to the partition. Thus, the silhouette value *silh* of a citizen cz_n within a sub-partition p_i was provided by comparing the average distance d_i between cz_n and all other citizens in p_i with the average distance q between cz_n and all sub-partitions in the neighbor partition p_j. *silh* was given by the following equation:

$$silh(cz_n) = \frac{q(cz_n) - d(cz_n)}{\max(d(cz_n), q(cz_n))}. \tag{5}$$

In addition to its ability to provide information about a single citizen quality of classification, the silhouette value can be extended to evaluate the entire clustering C. The average silhouette width $silh(p_i)$ of sub-partition p_i is the average silhouette value for all members within the sub-partition. It is defined as follows:

$$silh(p_i) = \frac{1}{|p_i|} \sum_{cz_n \in p_i} silh(cz_n). \tag{6}$$

In fact, for each sub-partition p_i, silhouette function returned the partition to which p_i belonged as well as the neighbor partition of p_i and the silhouette width $s(i)$ of the sub-partition.

In addition, silhouette plot provided the following information: "*cluster*," "*neighbor*" and "*sil-width*." Figure 6 reveals the resulting silhouette plot of the used real network.

As shown in Figure 6, the introduced mixed hierarchical citizens clustering method performs almost perfectly and generally outperforms the quality of baselines approaches. Obviously, the closeness of citizens in its own sub-partition is higher than its closeness in others partitions, indicating that the social network citizens with similar degree of danger are well placed in their detected communities. Moreover, a strong structure of citizens groups was found by applying the introduced citizen classification method with an average silhouette width = 0.83. Besides, most sub-partitions were well classified and belonged to the appropriate partition. However, referring to the average silhouette width of Nave Bayes and SVM classifier (0.53 for Nave Bayes and 0.52 for SVM), we remark the existence of a reasonable citizens' community structure. The accuracy between citizens and sub-partitions found using Verma' technique is weak (with silhouette width = 0.45), demonstrating that social network citizens were not placed on their adequate sub-partition and they might really belong to other partitions.

4.2.2 Clustering quality based on DBI. DBI is a well-known criterion for internal evaluation of clustering results. It assigns the best score to methods providing clusters with

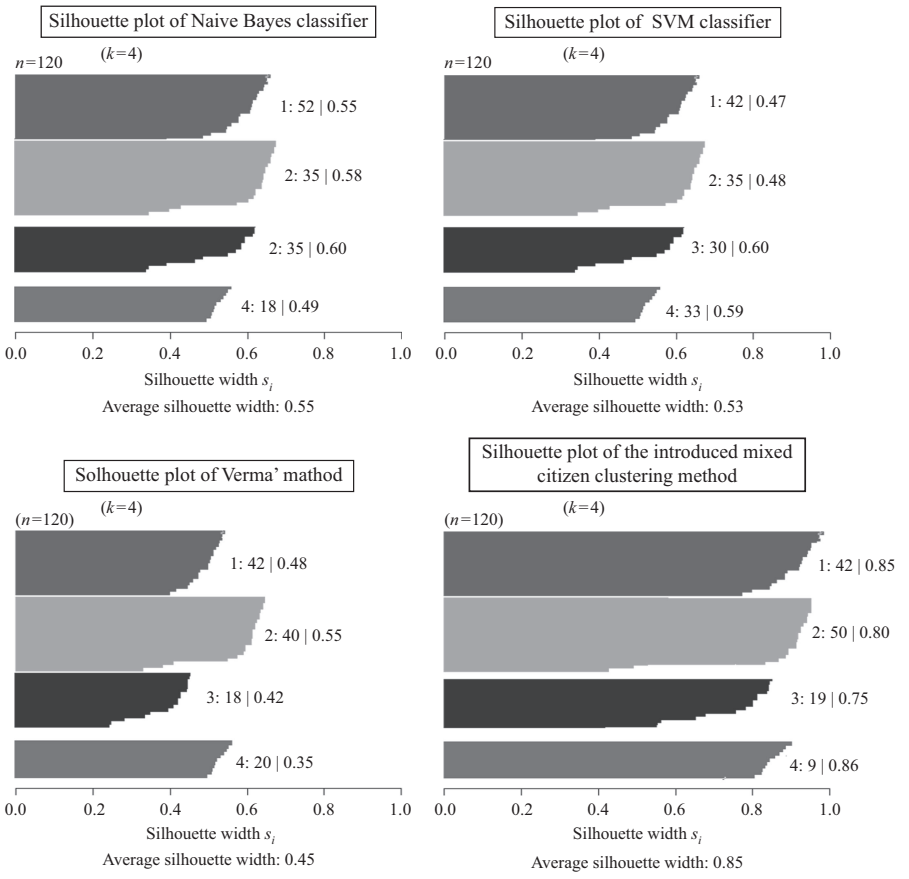


Figure 6.
Comparison of the clustering quality in terms of silhouette values

high similarity within a cluster and low similarity between clusters. The *DBI* was computed as follows:

$$DBI = \frac{1}{k}\sum_{c_i \in C} \max^{c_i \neq c_j}\left(\frac{\sigma_i + \sigma_j}{\overline{d(c_i, c_j)}}\right), \tag{7}$$

where k is the number of cluster c_i and σ represents the centroid of cluster. As it is observed from Equation (7), *DBI* relates the average distance of the elements of each cluster from their respective centroid to the distance of the centroid of the two clusters.

In Figure 7, we compare the performance (in terms of *DBI*) of baselines approaches to that of the introduced citizens clustering method vs the different cluster sizes.
Considering Figure 7, which plots *DBI* for CrisisLexT26 data set against the number of clusters, the introduced mixed hierarchical citizens clustering method displays better clustering, especially as the cluster number grows (with Davies index = 0.22 and cluster number = 8). Besides, our technique displays better clustering quality than the baselines approaches. It is also observed that the difference between the clustering quality of SVM and that of Naves Bayes classifiers is negligible displaying the worst classification performance. Besides, citizens' partitions provided by Verma' method are poorly separated.

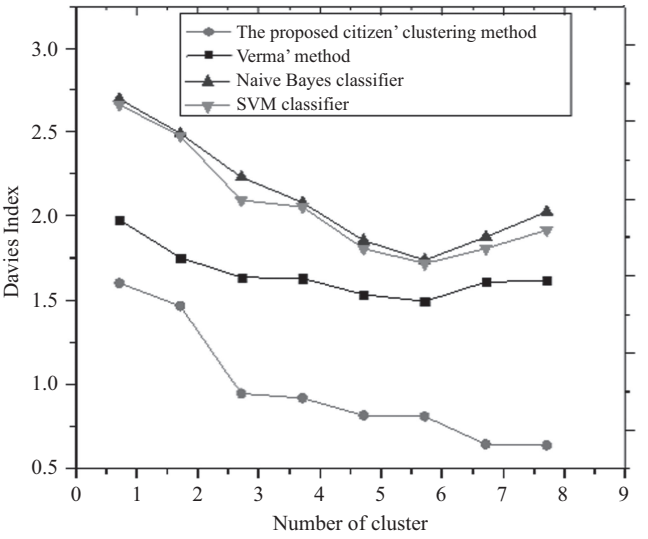


Figure 7.
Clustering quality comparison in terms of Davies index values

Obviously, the introduced method produced partitions with low intra-cluster distance and high inter-cluster distance, and provided the lowest Davies index values and the most similar clusters with high intra-cluster similarity and low inter-cluster similarity.

5. Visualization

Results provided by the developed mixed citizen' clustering technique for CrisisLexT26 data set were visualized exploring power BI dashboards for visual analytics. Obviously, visualization is effective for breaking news, as well as for quickly imparting new information like the location of a crisis and the number of casualties and for feature stories. Thus, a journalist can go deeper into a topic and offer a new perspective to help the reader see the news event in a completely new way. As shown in Figure 8, power

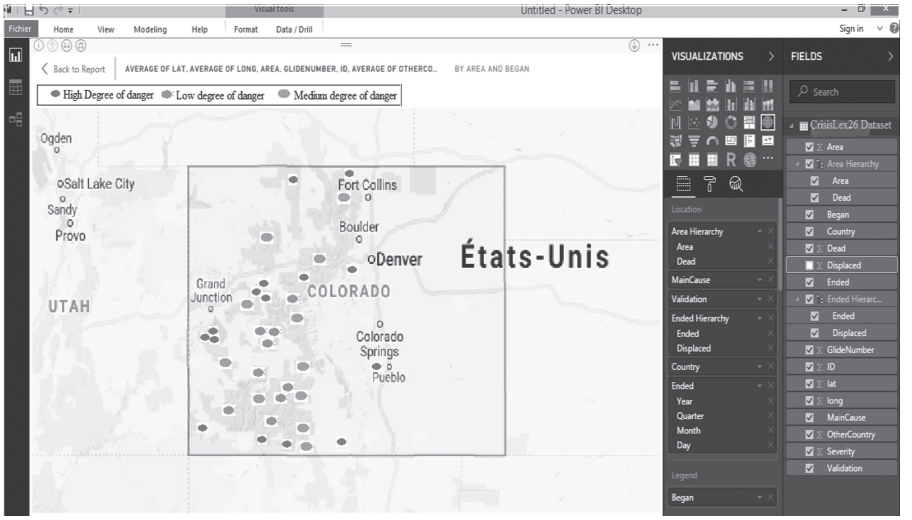


Figure 8.
Citizens clustering visualization using power BI maps

BI maps allows journalists to analyze data using a simple generated dashboard which indicates the most infected areas and delivers powerful awareness which can assists emergency operations.

In fact, "3D" maps for disaster management systems can accelerate assessment including huge amount of detailed information about the location and the degree of citizens' danger per region.

6. Discussion and conclusion

The important issues discussed in this paper are: first, the detection and the extraction of natural hazard events from social news contents in order to help journalists identify eyewitnesses and contextualize the event news, which increases the potential for the social media content to be discovered, used, shared and integrated in journalist reports. Second is the introduction of fuzzy processing generating the degree of danger for each extract news event as veracity indicator. Fuzzy indicators would be useful for journalists to summarize the overall quality of the provenance trace and to show whether content comes from a reputable source. Third is the development of efficient process to assess the quality of citizens' clustering by combining divisive and agglomerative hierarchical clustering and focusing on both noise and burst news detection. And fourth is facilitating the sharing and integration of news in digital journalism, by applying a visualization process. In fact, classified citizens were visualized exploring power BI dashboards for visual analytics. Moreover, the combination of the strengths of bottom-up hierarchical clustering with those of top-down clustering produced a meaningful mixed hierarchical citizens' community structure. While an agglomerative hierarchical method was good at filtering non-relevant news event from small clusters but not a large one, the strengths of divisive technique were reversed, which requires more complex analysis. However, the proposed mixed hierarchical citizens clustering model provided good results for noise and burst citizens' detection. The evaluation of the development of citizens' clustering approach was conducted on both a public benchmark for the problem of topic detection in streams of social content and real networks constructed by "CrisisLex T26" data from Twitter. In fact, the introduced mixed hierarchical citizens clustering provided low energy function of the elements of SNOW Data Challenge, indicating good clustering quality. This high clustering performance can result from the fact that the detected citizens communities were very dependent on the initial citizens community structure relying on the detection of the optimal initial structure (Toujani and Akaichi, 2017). Besides, because the introduced mixed process allowed adjusting the number of clusters, the hierarchical classification process was more flexible for noise and burst detection. Indeed, the alternative combination of the aggregation and the decomposition reduced the cost function of the clustering. Hence, if the cluster was to be divided into two sub-clusters, the cost of the clustering would be reduced after each divisive hierarchical level in the introduced method. Consequently, the evaluation of the detected citizens groups by our method relying on *CEC* provided the lowest cost function compared with baselines methods. In addition, external validation measures used in this study, namely, *CEC* and Precision-Recall applied the on SNOW benchmark, showed high dependency detected citizen' community.

Indeed, the main objective of our work is to provide cohesive and separate citizens partitions by referring to the internal criteria in order to measure the effectiveness of a specific clustering structure without reference to external information. In this research work, we employed silhouette and *DBI* to evaluate the partitions of the introduced method based on events news extracted from CrisisLexT26 data set. Davies–Bouldin values proved that

the proposed mixed hierarchical citizens' technique allowed obtaining partitions with high compact and close members. Moreover, the average silhouette width for a cluster indicated that the introduced method generated both well-separated and cohesive citizens partitions with close degree of danger.

Obviously, experimental results revealed good clustering quality of the proposed method for both evaluations indices, namely, external criteria and internal criteria.

The knowledge, experience and intuition of journalists are too fine to be replaced by current technologies. Indeed, our goal is to create tools that mediate the inhumane amounts of data and content created on a daily basis. We may also conclude that detecting hierarchical communities can be efficiently used to know the hidden features and to have a clear idea about the network structure. Nevertheless, to achieve such task, we must focus on the change of social news over time. Thus, in our future research work, we intend to concentrate on the evolution of citizens' community structure.

Notes

1. www.cis.uni-muenchen.de/schmid/tools/TreeTagger/
2. CrisisLexT26 data set, available at: www.crisislex.org/data-collections.html#CrisisLexT26

References

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R. and Tao., K. (2012), "Twitcident: fighting fire with information from social web streams", *Proceedings of the 21st International Conference on World Wide Web ACM*, pp. 305-308.

Boididou, C., Middleton, S.E., Jin, Z., Papadopoulos, S., Dang-Nguyen, D.T., Boato, G. and Kompatsiaris, Y. (2018), "Verifying information with multimedia content on Twitter", *Multimedia Tools and Applications*, Vol. 77 No. 12, pp. 15545-15571.

Brandtzaeg, P.B., Lüders, M., Spangenberg, J., Rath-Wiggins, L. and Følstad, A. (2016), "Emerging journalistic verification practices concerning social media", *Journalism Practice*, Vol. 10 No. 3, pp. 323-342.

Deborah, S., Chung, S.N. and Yamamoto, M. (2017), "Conceptualizing citizen journalism: US news editorsâ^{TM} views", *Journalism*, 1464884916686596, available at: https://doi.org/10.1177/SAGE-JOURNALS-UPDATE-POLICY

Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Hoi, G.W.S. and Zubiaga, A. (2017), "Semeval-2017 task 8: rumoureval: determining rumour veracity and support for rumours", arXiv preprint arXiv:1704.05972.

Diakopoulos, N., Choudhury, M.D. and Naaman, M. (2012), "Finding and assessing social media information sources in the context of journalism", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems ACM*, pp. 2451-2460.

Dou, W., Wang, X., Ribarsky, W. and Zhou, M. (2012), "Event detection in social media data", *IEEE VisWeek Workshop on Interactive Visual Text Analytics-Task Driven Analytics of Social Media Content*, pp. 971-980.

DouEnayet, O. and El-Beltagy, S.R. (2017), "Niletmrg at semeval-2017 task 8: determining rumour and veracity support for rumours on Twitter", *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 470-474.

Garg, M. and Kumar, M. (2016), "Review on event detection techniques in social multimedia", *Online Information Review*, Vol. 40 No. 3, pp. 347-361.

Khan, M.A.H., Bollegala, D., Liu, G. and Sezaki, K. (2013), "Multi-tweet summarization of real-time events", *IEEE 2013 International Conference on Social Computing (So-cialCom)*, pp. 128-133.

Lin, C.-Y., Li, T.-Y. and Chen, P. (2016), "An information visualization system to assist news topics exploration with social media", *Proceedings of the 7th 2016 International Conference on Social Media & Society ACM*, p. 23.

Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S. and Miller, R.C. (2011), "Twitinfo: aggregating and visualizing microblogs for event exploration", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems ACM*, Taylor & Francis Group, Boca Raton, FL, pp. 227-236.

Martinez, W.L., Martinez, A.R. and Solka, J. (2017), *Exploratory Data Analysis with MATLAB*, Chapman and Hall/CRC.

Meladianos, P., Xypolopoulos, C., Nikolentzos, G. and Vazirgiannis, M. (2018), "An optimization approach for sub-event detection and summarization in Twitter", *European Conference on Information Retrieval Springer*, pp. 481-493.

Moshtaghi, M., Bezdek, J.C., Erfani, S.M., Leckie, C. and Bailey, J. (2018), "Online cluster validity indices for streaming data", arXiv preprint arXiv:1801.02937.

Olteanu, A., Castillo, C., Diaz, F. and Vieweg, S. (2014), "Crisislex: a lexicon for collecting and filtering microblogged communications in crises", *ICWSM, June*.

Ozdikis, O., Senkul, P. and Oguztuzun, H. (2012), "Semantic expansion of tweet contents for enhanced event detection in Twitter", *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society*, pp. 20-24.

Papadopoulos, S., Corney, D. and Aiello, L.M. (2014), "Snow 2014 data challenge: assessing the performance of news topic detection methods in social media", SNOW-DC@ WWW, pp 1-8.

Phuvipadawat, S. and Murata, T. (2010), "Breaking news detection and tracking in Twitter", *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Vol. 3, pp. 120-123.

Phuvipadawat, S. and Murata, T. (2011), "Detecting a multi-level content similarity from microblogs based on community structures and named entities", *Journal of Emerging Technologies in Web Intelligence*, Vol. 3 No. 1, pp. 11-19.

Rui, Y. (2012), "A hierarchical clustering and validity index for mixed data".

Sayyadi, H., Hurst, M. and Maykov, A. (2009), "Event detection and tracking in social streams", *Icwsm*, Association for the Advancement of Artificial Intelligence.

Shi, L.-L., Liu, L., Wu, Y., Jiang, L. and Hardy, J. (2017), "Event detection and user interest discovering in social media data streams", *IEEE Access*, Vol. 5, pp. 20953-20964, doi: 10.1109/ACCESS. 2017.2675839.

Spurek, P. (2017), "Split-and-merge tweak in cross entropy clustering", *Computer Information Systems and Industrial Management: 16th IFIP TC8 International Conference, CISIM 2017, Proceedings Springer*, Vol. 10244, *Bialystok, June 16–18*, p. 193.

Stephens-Davidowitz, S. and Pinker, S. (2017), "Everybody lies: big data, new data, and what the internet can tell us about who we really are", HarperCollins, New York, NY.

Stowe, K., Paul, M.J., Palmer, M., Palen, L. and Anderson, K. (2016), "Identifying and categorizing disaster-related tweets", *Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media*, pp. 1-6.

Toujani, R. and Akaichi, J. (2017), "Optimal initial partitionning for high quality hybrid hierarchical community detection in social networks", *Proceedings of the International Conference on Control, Decision and Information Technologies(CoDIT'17)*, April 5-7, Barcelona.

Verma, S., Vieweg, S., Corvey, W.J., Palen, L., Martin, J.H., Palmer, M., Schram, A. and Anderson, K.M. (2011), "Natural language processing to the rescue? Extracting 'situational awareness' tweets during mass emergency", *ICWSM*, Association for the Advancement of Artificial Intelligence, Barcelona, pp. 385-392.

Wei, H., Sankaranarayanan, J. and Samet, H. (2017), "Finding and tracking local Twitter users for news detection.

Wei, X., Zhu, F., Jiang, J., Lim, E.-P. and Wang, K. (2016), "Topicsketch: realtime bursty topic detection from Twitter", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28 No. 8, pp. 2216-2229.

Zadeh, L.A. (1999), "Fuzzy sets as a basis for a theory of possibility", *Fuzzy Sets and Systems*, Vol. 100, pp. 9-34, available at: https://doi.org/10.1016/0898-1221(83)90013-5

Zubiaga, A., Ji, H. and Knight, K. (2013), "Curating and contextualizing Twitter stories to assist with social newsgathering", *Proceedings of the 2013 International Conference on Intelligent User Interfaces ACM*, pp. 213-224.

Corresponding author
Radhia Toujani can be contacted at: toujaniradia@gmail.com

An exploratory approach to the computational quantification of journalistic values

Sujin Choi
Kyung Hee University, Seoul, Republic of Korea

Received 15 March 2018
Revised 12 July 2018
4 October 2018
Accepted 5 October 2018

Abstract

Purpose – News algorithms not only help the authors to efficiently navigate the sea of available information, but also frame information in ways that influence public discourse and citizenship. Indeed, the likelihood that readers will be exposed to and read given news articles is structured into news algorithms. Thus, ensuring that news algorithms uphold journalistic values is crucial. In this regard, the purpose of this paper is to quantify journalistic values to make them readable by algorithms through taking an exploratory approach to a question that has not been previously investigated.

Design/methodology/approach – The author matched the textual indices (extracted from natural language processing/automated content analysis) with human conceptions of journalistic values (derived from survey analysis) by implementing partial least squares path modeling.

Findings – The results suggest that the numbers of words or quotes news articles contain have a strong association with the survey respondent assessments of their balance, diversity, importance and factuality. Linguistic polarization was an inverse indicator of respondents' perception of balance, diversity and importance. While linguistic intensity was useful for gauging respondents' perception of sensationalism, it was an ineffective indicator of importance and factuality. The numbers of adverbs and adjectives were useful for estimating respondents' perceptions of factuality and sensationalism. In addition, the greater numbers of quotes, pair quotes and exclamation/question marks in news headlines were associated with respondents' perception of lower journalistic values. The author also found that the assessment of journalistic values influences the perception of news credibility.

Research limitations/implications – This study has implications for computational journalism, credibility research and news algorithm development.

Originality/value – It represents the first attempt to quantify human conceptions of journalistic values with textual indices.

Keywords Digital journalism, Computational journalism, Credibility, Journalistic value, News algorithm

Paper type Research paper

The current information environment is subject to the attention economy (Lanham, 2006): while our cognitive resources (i.e. attention) are limited, the information we can access is not. In this circumstance, which information or news is produced is less important than which information or news gains our attention is (Choi and Kim, 2017; Lotan, 2014). News algorithms[1] help us allocate our attention to particular news or information through the process of classification, association and prioritization (Diakopoulos, 2015). In brief, algorithms classify news articles into certain categories (e.g. politics, business, technology, science, sports, food and international affairs), form clusters on the basis of issue similarity within particular categories and then prioritize the news articles included in each particular cluster. This process ranks certain news articles higher than others and places them at the top of the screen, where people's attention usually remains (Nielsen, 2006; Sherman, 2005). Therefore, the chance of news articles being exposed to and read by users is structured into the news algorithms (Ananny, 2016; Gillespie, 2014).

This work was supported by a grant from Kyung Hee University in 2018 (KHU-20180926). In addition, the author is grateful to the Korea Press Foundation's News Trust Committee, of which the author was a founding member, for helping inspire this study.

This paper forms part of a special section "Social media mining for journalism".

Online Information Review
Vol. 43 No. 1, 2019
pp. 133-148

1468-4527
DOI 10.1108/OIR-03-2018-0090

News algorithms' article classification, association and prioritization processes resemble the framing process traditionally used by journalists and legacy news organizations. Framing – defined as a process of "persistent selection, emphasis, and exclusion" (Gitlin, 1980, p. 7) of perceived reality – influences audience perception, organization and interpretation of numerous pieces of information (Goffman, 1974). In controlling news articles' potential exposure to online news users, news algorithms influence said users' perceptions of reality – a process that has been labeled "algorithmic reality construction" (Just and Latzer, 2017, p. 246). In addition to potentially affecting people's perceptions of reality, such reality construction may also affect democratic decision-making processes (Bucher, 2017; Gillespie, 2014; Segev, 2017). In this regard, news algorithms not only help us efficiently navigate the sea of available information, but also frame information in ways that influence public discourse and citizenship.

This raises several important questions. How can we ensure that these publicly important news algorithms (which select, rank and display news articles) uphold journalistic values? Specifically, in what ways can we design these algorithms to understand the notion of journalistic values and how can we quantify journalistic values? Except for the developers and owners of news algorithms, no one currently has a clear answer regarding the values or criteria news algorithms use; companies that own news algorithms are not yet legally obliged to disclose this information[2]. However, increasing academic and public demand for algorithmic transparency/accountability (e.g. Ananny, 2016; Diakopoulos, 2015; Diakopoulos and Koliska, 2017; Dörr and Hollnbuchner, 2017; Just and Latzer, 2017; Lotan, 2014; Oh, 2016; Oh and Kim, 2016; Stark and Diakopoulos, 2016) may lead to more open discussion of the principles on which news algorithms are based, which could help ensure that developers orient their algorithms to realize journalistic values. Concern regarding covert news algorithms is already part of the public consciousness, as people wonder: "what happens when information providers no longer care about truth or accuracy?" (Sambrook, 2012, p. 10).

Furthermore, the negative side effects of the contemporary digital news environment highlight the need for news algorithms that uphold journalistic values. The current digital news environment has been described as a "content farm" or a "digital sweatshop" (Bakker, 2012). To survive the so-called "news cyclone," online news producers constantly publish articles that have not gone through sufficient fact checking (Klinenberg, 2005). The repetitive news phenomenon (Choi and Kim, 2017) and churnalism (Davies, 2009) also prevail. News algorithms contribute to these phenomena by privileging more recent and more search-word relevant news articles. To increase algorithmic "choosability" and thereby bolster "clickability," online news producers tend to churn out press releases, become more repetitive, disregard fact checking processes and make stories more provocative. This series of phenomena has fueled the decline of news credibility.

Acknowledging this context, this study took an exploratory approach to the quantification of journalistic values (balance, variety, importance, factuality, readability and sensationalism), based on the assumption that news articles that reflect journalistic values are credible. Our quantification of journalistic values had two levels: the quantification of human conceptions of journalistic values and the quantification of these conceptions into natural language processing (NLP) indices that computers understand. We reasoned that, for instance, the identification of news articles that people perceive as balanced must precede the identification of the NLP indices that reflect human conceptions of balance. Furthermore, people may not rely on a single journalistic value to assess the credibility of specific news articles. Thus, we found it necessary to address the question of which journalistic values explain news credibility.

In order to answer the questions above, we used the following methods: a survey of human perception of journalistic values, the NLP on R platform, an automated content

analysis using an emotion lexicon, and partial least squares (PLS) path modeling between quantified indices and human perceptions of journalistic values using SmartPLS. Our findings have theoretical implications for computational journalism and credibility research and practical implications for news algorithm development.

News credibility and journalistic values

Journalism researchers have engaged in extensive discussions of journalistic values. Influenced by social and technological changes, researchers have emphasized different journalistic values at different times (Dörr and Hollnbuchner, 2017), but certain journalistic values have remained consistently important to ensuring that journalism retains its original form. The values traditionally agreed upon by both academics and journalists include accuracy, fairness, objectivity, impartiality, truthfulness and sincerity (ASNE[3]; BBC[4]; CBC[5]; Deuze and Yeshua, 2001; Kovach and Rosenstiel, 2007; Sambrook, 2012; Singer, 2010; Steele, 2008; Van Der Wurff and Schönbach, 2011). Although researchers have debated the practicality of certain values – for example, objectivity and impartiality (Cohen-Almagor, 2008; Singer, 2010) – these values have been generally accepted as norms that professional journalists must fulfill.

A value that currently receives considerable attention, diversity, is a more practical value than impartiality or objectivity (Sambrook, 2012). The BBC emphasizes diversity by covering a wide range of views. As a criterion for quality news, Pew's reporting index[6] also considers a mix of viewpoints and a multitude of stakeholders. The value of diversity has become even more important in today's politically fragmented and polarized online discourse (Sambrook, 2012). Addressing diverse opinions within a single news report may help attenuate the selective exposure phenomenon (i.e. consuming news that supports one's political predispositions). Online news editors have also identified content relevance and good writing as important criteria for online news content (Gladney *et al.*, 2007) – news content closely related to audiences, addressing local issues and containing elites/public figures has strong relevance and importance, and its readability depends on writing quality. The recurring discussion of "good stories" vs "good journalism" highlights the issue of sensationalism. For news producers, good stories typically concern terror, war, rape and violence, which draw public attention; however, in many cases, such stories fail to fulfill the standards of good journalism (Cohen-Almagor, 2008). The more sensational the news report, the less likely it is to be of good quality.

The fulfillment of certain journalistic values has been presumed to consequently enhance the trustworthiness or credibility of news articles (Hayes *et al.*, 2007). However, this presumption widely accepted in journalism studies has not been empirically tested. We do not yet have a clear understanding of the associations between journalistic values and individuals' perceptions of the credibility of the news stories they read.

Many previous studies have investigated the credibility of sources and the effects of source-credibility on persuasion (Wilson and Sherrell, 1993) or the credibility of mediums and its determinants (Johnson and Kaye, 2014), but few studies have examined the credibility of messages *per se* (Sundar, 1999). Among the few studies that have specifically addressed the credibility of news, Sundar (1999) identified several measures that influence people's perceptions of online news stories. These measures include "biased, fair, objective, clear, coherent, comprehensive, concise, well-written, important, relevant, and timely." Participants in Cassidy (2007) evaluated online news credibility based on believability, fairness, accuracy, and comprehensiveness. In addition to the aforementioned criteria, researchers have identified local coverage, a lack of sensationalism, interest/usefulness to readers, depiction of subjects in favorable lights and mechanical/grammatical accuracy as important standards that people use to assess news content excellence (Gladney, 1996; Sallot *et al.*, 1998).

When it comes to the credibility of the press or journalism more broadly, research has identified measures including "being objective, covering stories that should be covered, helping people, getting information to the public quickly, providing analyses and interpretation of complex problems, verifying facts, giving ordinary people a chance to express their views, and being the watchdog for the public" as the "professional tenets of good journalism" (Holton *et al.*, 2013). Regarding press credibility, researchers have also investigated the extent to which people regard the press as "objective, fair, accurate, non-sensational, considerate of the readers' interest, not seeking commercial profits, care about the public interest, and can be trusted" (Choi and Kim, 2017).

The discussions above reveal the potential link between news credibility criteria and journalistic values. However, we still do not know how individual perceptions of journalistic values influence individual evaluations of news credibility. For instance, if people perceive a news article as balanced, would it lead them to assess said article as highly credible? Would individuals' perceptions of news articles' balance or factuality have a greater influence on their evaluations of the articles' credibility? These questions remain unexplored.

We derived six journalistic values – readability, diversity, importance, factuality, balance and sensationalism – from previous discussions of the journalistic values that contribute to news trustworthiness and the criteria that people use to assess news credibility. Past studies have considered other values including relevance, depth and transparency (Gladney *et al.*, 2007; Hayes *et al.*, 2007; Van Der Wurff and Schönbach, 2011), but these have been far less frequently addressed than the six aforementioned values. Furthermore, we chose the six journalistic values that have relatively lower likelihoods of misinterpretation and stronger conceptual consensus in a general sense. For instance, the relevance of news stories may vary considerably from individual to individual and issue to issue – news articles about the job market, for example, will have greater utility and relevance for the unemployed than for others; the judgment of news story depth may be easily confounded with news article length; and, the question of news article transparency may be conflated with the question of factuality – although transparency generally concerns "how" (i.e. the route that the information or clue is obtained) (Deuze, 2005; Hayes *et al.*, 2007; Karlsson, 2010) and factuality concerns "what" (i.e. the fact-based subject of a news article and its accuracy) (Sallot *et al.*, 1998), this conceptual distinction is subtle and hard to perceive, and therefore we chose factuality because it is conceptually clearer than transparency.

Many people, both journalists and non-journalists, consider credibility an important value for news to pursue (Cassidy, 2007; Mitchell *et al.*, 2016). In the digital news environment where inaccurate and fake information spreads easily, assessments of news credibility are particularly important. In this regard, the factors that prompt people to assess news messages as credible require further investigation. We attempt to identify the journalistic values perceived by individuals that best explain their perceptions of news credibility. We also aim to quantify their perceived journalistic values in ways that the machines that select, prioritize and display news articles that individuals consume in their daily lives could understand.

Journalistic values and their quantification

As Figure 1 shows, the aforementioned six values can be quantified into 14 indices created using NLP and automated content analysis; we elaborate on these indices in the Methods section below. We attempt to determine which indices better quantify which values. To our knowledge, almost no research has been done on the associations between textual indices and journalistic values. Acknowledging the limits of prior knowledge, we made the following tentative presumptions.

The numbers of words or quotes may relate to balance, diversity, importance and factuality, since, to be perceive as balanced, inclusive of diverse perspectives, focused

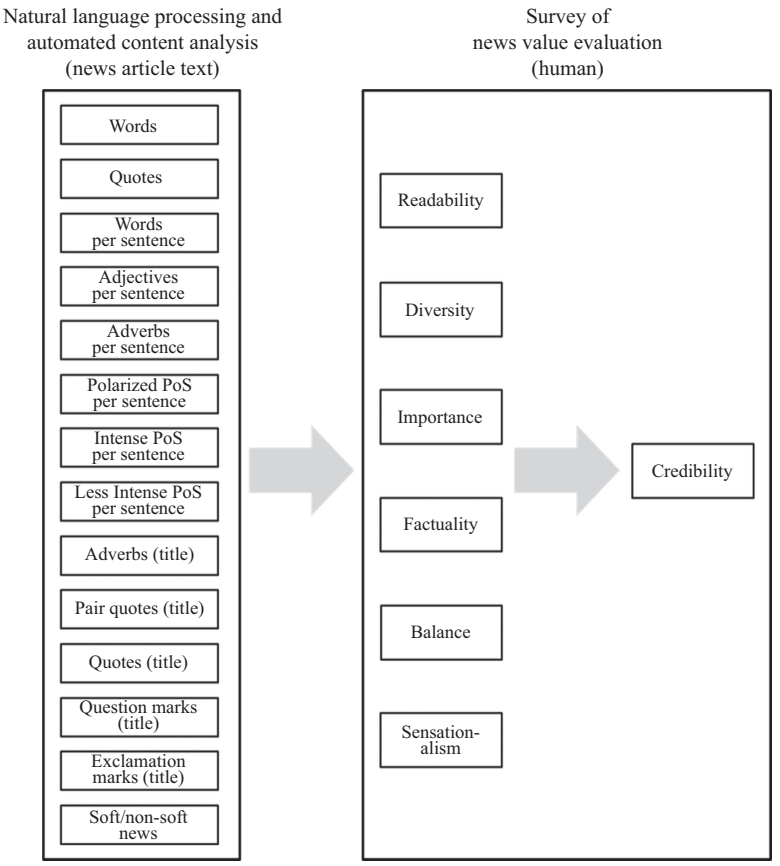


Figure 1.
Research model

on important issues, and based on facts, news articles must be of a certain length and contain a certain number of sources. The number of words per sentence may relate to readability – the more words appear per sentence, the more complex and less readable the news article might be.

The numbers of adjectives and adverbs per sentence may be associated with factuality and sensationalism. The more adjectives and adverbs used per sentence, the more descriptive, vivid and verbose the writing becomes. Therefore, in a sense, more flowery words may make news articles more sensational and less factual. On the other hand, more adjectives and adverbs may contribute to better descriptions of incidents and thus make people perceive news articles as more factual.

Linguistic polarity may relate to balance and diversity. If news articles are emotionally polarized (positively or negatively), they may not be balanced and may not treat diverse opinions fairly. Linguistic intensity may relate to importance, factuality, readability and sensationalism. The use of strong and intense words may accentuate the importance of the incidents the news articles are covering. It may also help readers more easily understand news articles' key points. However, the more intense words news articles use, the more likely people may be to perceive them as sensational, provocative and less factual. People may regard straight news that simply explains who, what, when, where and how in plain language as more factual than news articles that include strong and highly charged words.

In addition to examining the words in the main text of news articles, we also considered the words in the headlines. News headlines become more important in the digital news environment, since headlines and a few introductory sentences are all online news users see before clicking. Headlines thus determine whether or not people read news articles – the critical role of headlines is the source of the so-called headline journalism. The textual indices extracted from news headlines include the numbers of adverbs, pair quotes, quotes, question marks and exclamation marks. These indices usually appear in a headline to distinguish it from all the others listed on the screen of a digital device. Therefore, they may positively relate to sensationalism and importance. In addition, pair quotes or quotes in headlines may signal the main text's diverse sources or factual basis – thus relating to diversity and factuality.

News type may also be associated with journalistic values. Soft news (entertainment, sports and social issues) cannot be assessed using the same criteria as non-soft news (politics, economics, business and international affairs). Non-soft news may be more balanced, more diverse, more important, more factual, less readable and less sensational than soft news. We accounted for this seeming difference by including the type of news in our research model.

In addition to the conceptual presumptions explained above, we conducted a preliminary analysis to examine the hidden relationships between textual indices and journalistic values. Since this was the first attempt to quantify journalistic values, we relied on both conceptual reasoning and statistical analysis to reveal possible relationships. We implemented a multivariate regression analysis by inserting six journalistic values as dependent variables and 14 textual indices as independent variables. This analysis, along with the conceptual reasoning above, provided us preliminary guidelines for building a research model. Details can be found in the Results section below.

Methods

Data set description

The data set included 1,000 news articles randomly selected from articles published online between January 1, 2015 and June 30, 2016. We gave these articles to eight graduate students majoring in Journalism. They read each article and responded to survey questions appended to each article. The questions asked respondents to evaluate on a ten-point scale the balance, diversity, importance, factuality, readability and sensationalism of each news article. Regarding readability, for instance, higher scores signal greater readability. After evaluating these journalistic values, respondents also assessed each news article's overall credibility on a ten-point scale.

We designed this survey to measure respondents' perceptions of the journalistic values, not to content-analyze the news articles into certain categories. The object of study is not the news articles *per se*, but people's perceptions of journalistic values reflected through evaluating journalistic values of each news article. Our ultimate goal is to quantify people's perceptions of journalistic values by regressing the journalistic value scores respondents assigned to the news articles (obtained via survey) on the news articles' text indices (obtained via NLP and automated content analysis).

Considering the immense amount of time and effort required to read 1,000 articles, we inevitably used a single indicator to measure journalistic values. We gave respondents one month to complete the survey. The News Trust Committee of the Korea Press Foundation[7] implemented and funded the data collection and survey. The author of this paper is a founding member of this Committee and participated in conducting the survey.

Acknowledging that innate criteria for judging journalistic values might differ from one respondent to the next, we calculated *z*-scores for each individual's responses. Table I shows the means and standard deviations for individuals' journalistic value scores. The *z*-scores enabled us to make comparisons among individuals' responses at the same level and create

	Respondent							
	1	2	3	4	5	6	7	8
Balance								
Mean	2.1	2.8	2.2	2.2	3.8	3.1	5.0	3.1
SD	1.5	1.2	2.2	1.8	1.5	2.3	0.6	1.3
Variety								
Mean	2.5	2.6	2.2	2.8	3.8	4.4	5.0	2.9
SD	1.7	0.9	2.2	1.7	1.7	1.8	1.3	1.0
Factuality								
Mean	5.7	6.4	7.4	6.7	5.8	6.8	5.9	4.7
SD	2.6	1.3	1.3	1.3	1.8	1.1	1.3	1.2
Importance								
Mean	3.1	3.1	3.3	2.1	3.4	4.1	4.5	3.5
SD	2.4	1.5	2.6	1.7	1.8	2.0	2.4	1.4
Readability								
Mean	5.4	6.7	7.8	6.7	5.9	6.9	6.4	5.4
SD	2.3	0.8	1.0	1.1	1.6	0.7	1.0	1.3
Sensationalism								
Mean	2.4	2.7	2.9	1.5	1.9	2.7	1.9	2.1
SD	1.8	1.1	2.8	1.2	1.4	1.9	2.1	1.2
Credibility								
Mean	4.2	4.3	4.7	4.0	4.5	5.1	4.9	4.1
SD	1.9	1.3	2.5	1.9	1.5	1.8	2.0	1.1

Note: Survey respondents assessed journalistic values on a ten-point scale

Table I. Means and standard deviations for each respondent's journalistic value scores

composite scores across respondents' evaluation scores. This process ultimately generated eight *z*-scores for each news article (since the study included eight respondents). We then averaged these *z*-scores for each article and also calculated their standard deviations.

The standard deviation of the *z*-scores for each article served as a cutoff point for the elimination of articles. Because the specificity of this survey demanded that respondents exert considerable effort (reading 1,000 articles), the number of respondents was exceptionally low, and thus we could not statistically assume the central limit theorem. To address this limitation, we only included news articles whose *z*-score standard deviations had absolute values within one. As a result, our final data set included 938 news articles.

Natural language processing

For NLP in Korean, we used KoNLP library on R 3.4.3 platform. Using SimplePos22 command that has 22 part-of-speech (PoS) taggers, we analyzed 938 news article texts and conducted the PoS tagging morphologically. The number of PoS by article text ranged from 51 to 629 ($M = 216.2$, SD = 108.8).

In addition, we calculated the numbers of words, sentences, adverbs, adjectives and conjunctions that appeared in the main texts of the news articles. Considering the importance of news headlines in digital user interfaces, we also calculated the numbers of quotes, pair quotes, question marks, exclamation marks and adverbs used in the headlines.

Automated content analysis

We conducted an automated content analysis using the Korean Sentiment Analysis Corpus (KOSAC) developed by Shin *et al.* (2012). Since the PoS tagger labels differ for the KoNLP

and KOSAC, we renamed the KOSAC PoS taggers to correspond to those of the KoNLP. We made this modification to conduct the automated content analysis on the *R* platform.

We then content-analyzed the news articles' PoS tagging results by matching them with the KOSAC's classification of linguistic polarity and intensity. According to Shin *et al.* (2012), the polarity and intensity indices indicate, respectively, the valence and the strength of private states (i.e. the writers' mental or emotional states such as opinions, beliefs, thoughts, feelings, emotions and judgments). In terms of polarity, the positive PoS includes words such as gratitude, longing and strength, and the negative PoS includes blame, conflict and coercion. We used the absolute value of the difference between positive PoS and negative PoS – higher scores indicate higher polarity, whereas scores closer to zero indicate greater balance. In terms of intensity, the high-level PoS includes words such as accusation, vituperation and corruption, and the lower level PoS includes words such as remodeling, reinforcement, comparison and still.

KOSAC provides the sub-category classification probability of each PoS. For instance, it shows that a given PoS has a probability of 0.8 of being classified as positive in terms of linguistic polarity. To ensure accuracy as opposed to coverage, in this study, we only conducted content analysis on PoS that had probabilities of one.

PLS path modeling

The abovementioned methodological procedure generated a complete data set composed of news article ID numbers in rows and the average *z*-scores for six journalistic values and credibility, the NLP results and the automated content analysis results in the columns. Multicollinearity did not exist among the variables, which all had VIF scores below 2.72. To make the survey practically viable, we used a single indicator for journalistic value variables, as explained in the Data Description section.

Considering that this study used non-normal data[8] and took an exploratory approach, we employed PLS path modeling. While traditional structural equation modeling is covariance-based, assumes multivariate normality and relies on confirmatory theoretical underpinnings, PLS is variance-based, distribution-free and suited for exploratory purposes or for non-causal predictions (Garson, 2016; Vinzi *et al.*, 2010; Wong, 2013). PLS can contribute to research in the early stages of theoretical development by testing exploratory models (Tsang, 2002). For this analysis, we used SmartPLS software (Ringle *et al.*, 2015).

Results

Descriptive statistics

Table II shows the means and standard deviations of the news article-related variables. The average news article in our data set has 13.3 sentences (or 191.2 words) and contains around three quotes. On average, each sentence contains 16 words, 0.7 adjectives and 0.6 adverbs. In terms of polarity, each sentence averages 0.4 PoS – either positively or negatively polarized. In terms of intensity, each sentence averages 0.1 PoS of high intensity and 1.2 PoS of lower intensity. Among news articles in the data set, 55 percent are soft news (entertainment, sports and social news), and the remaining articles are non-soft news (political, economic, international affairs and local news).

Journalistic value and news credibility

The fit index of the PLS path model, the standardized root mean square residual, suggests an acceptable model fit (0.097). The adjusted R^2 value reveals that the model explains 89 percent of the variance in credibility, the endogenous dependent variable (see Table III).

While news article balance is not a significant predictor of news credibility, diversity, factuality, importance, readability and sensationalism all predict news credibility. The more

	Mean	SD
Sentences	13.26	8.45
Words	191.21	95.68
Quotes	2.90	3.27
Words per sentence	15.84	4.99
Adjectives per sentence	0.73	0.44
Adverbs per sentence	0.56	0.37
Polarized PoS per sentence	0.36	0.33
Intense PoS per sentence	0.12	0.15
Less intense PoS per sentence	1.17	0.57
Adverbs (headline)	0.25	0.53
Pair quotes (headline)	0.53	0.91
Quotes (headline)	0.80	1.18
Question marks (headline)	0.12	0.35
Exclamation marks (headline)	0.04	0.23
Soft/non-soft news[a]		
Soft news	517	
Non-soft news	421	

Notes: [a]Frequency information is provided for the dummy variable. This variable's value of zero indicates soft news, and its value of one indicates non-soft news

Table II.
Descriptive statistics

diverse, the more factual, the more important, the more readable and the less sensational respondents assess the news articles to be, the more likely they are to perceive them as credible. Among these factors, importance has the largest explanatory power, followed by diversity, factuality, sensationalism and readability.

Journalistic value quantification

Regarding balance, we found the numbers of words, quotes and polarized PoS per sentence and the classification of soft/non-soft news to be statistically significant (see Table III). The more words and quotes and the fewer polarized PoS that appear in the news articles, the less likely respondents are to perceive them as balanced. We also found that respondents perceive non-soft news as more balanced than soft news. These variables account for 21 percent of the variance in balance.

Diversity is predicted by the numbers of words, quotes and polarized PoS per sentence. The more words and quotes and the fewer polarized PoS news articles include, the more likely respondents are to perceive them as diverse. In addition, respondents assess news articles with headlines containing pair quotes or quotes as less diverse. We also found that respondents believe non-soft news better reflects the diversity value than soft news. Together, these variables explain 42 percent of the variance in diversity.

The numbers of words, quotes, intense PoS per sentence and less intense PoS per sentence show a positive association with importance. Interestingly, both intense and less intense PoS appear frequently in the news articles that respondents believe realized the importance value. The number of polarized PoS per sentence and the numbers of quotes, exclamation marks and question marks that appear in the news headline have a negative relationship with perceived importance. Respondents regard non-soft news as better at realizing the importance value than soft news. All these variables explain 27 percent of the variance in importance.

Our findings show that perceived factuality is positively associated with the numbers of words, intense PoS per sentence, and less intense PoS per sentence. As with importance, both intense and less intense PoSs have statistically significant relationships with factuality.

	Standardized path coeff.
Balance → Credibility	0.02
Words → Balance	0.33***
Quotes → Balance	0.16***
Polarized PoS per sentence → Balance	−0.10***
Soft/non-soft news → Balance	0.15***
Diversity → Credibility	0.36***
Words → Diversity	0.52***
Quotes → Diversity	0.21***
Polarized PoS per sentence → Diversity	−0.05*
Pair quotes (headline) → Diversity	−0.09**
Quotes (headline) → Diversity	−0.07**
Soft/non-soft news → Diversity	0.17***
Importance → Credibility	0.43***
Words → Importance	0.26***
Quotes → Importance	0.07*
Polarized PoS per sentence → Importance	−0.12***
Intense PoS per sentence → Importance	0.07*
Less intense PoS per sentence → Importance	0.08*
Quotes (headline) → Importance	−0.18***
Pair quotes (headline) → Importance	−0.01
Exclamation marks (headline) → Importance	−0.12***
Question marks (headline) → Importance	−0.13***
Soft/non-soft news → Importance	0.21***
Factuality → Credibility	0.21***
Words → Factuality	0.15***
Adjectives per sentence → Factuality	−0.14***
Adverbs per sentence → Factuality	−0.14***
Intense PoS per sentence → Factuality	0.09**
Less intense PoS per sentence → Factuality	0.15***
Quotes (headline) → Factuality	−0.10**
Soft/non-soft news → Factuality	0.21***
Readability → Credibility	0.09***
Words per sentence → Readability	0.06
Intense PoS per sentence → Readability	0.05
Soft/Non-soft news → Readability	0.13***
Sensationalism → Credibility	−0.16***
Adjectives per sentence → Sensationalism	0.09***
Adverbs per sentence → Sensationalism	0.07*
Intense PoS per sentence → Sensationalism	0.08*
Less intense PoS per sentence → Sensationalism	−0.15***
Adverbs (headline) → Sensationalism	0.07
Quotes (headline) → Sensationalism	0.14***
Pair quotes (headline) → Sensationalism	0.15***
Exclamation marks (headline) → Sensationalism	0.07
Question marks (headline) → Sensationalism	0.16***
Soft/non-soft news → Sensationalism	−0.10***
Overall adjusted R^2	
Credibility	0.89
Balance	0.21
Diversity	0.42
Importance	0.27
Factuality	0.17
Readability	0.03
Sensationalism	0.17

Table III.
PLS path model result

Notes: Regarding soft/non-soft news variables, we coded soft news as 0 and non-soft news as 1. p-value is calculated by bootstrapping (two-tailed test). *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

We found a negative association between factuality and the numbers of adjectives per sentence, adverbs per sentence, and quotes in news headlines. Respondents perceive non-soft news as more factual than soft news. Collectively, these variables account for 17 percent of the variance in factuality.

Regarding readability, only the classification of soft/non-soft news is statistically significant. Respondents perceive non-soft news as more readable than soft news. Contrary to our expectations, the verbosity and the intensity of words do not affect readability. These variables account for 3 percent of the variance in readability.

Sensationalism is the only variable with a negative relationship to credibility. The more adjectives, adverbs and intense PoS that appear in the news articles, the more likely respondents are to perceive the articles as sensational and provocative. Notably, contrary to our findings regarding importance and factuality, we found a negative association between the number of less intense PoS per sentence and sensationalism. We also found that the numbers of quotes, pair quotes and question marks in the headlines serve as significant indicators of perceived sensationalism. As expected, respondents perceive non-soft news as less sensational than soft news. These variables explain 17 percent of the variance in sensationalism.

As mentioned in the Method section, the aforementioned findings in regards to journalistic values are based on the survey responses of a small number of journalism graduate students. Thus, caution should be taken to generalize these results beyond the group of study.

Discussion and conclusion

Social science research has devoted an increasing amount of attention to algorithmic transparency and accountability. However, the current agenda lacks a clear sense of the values on which algorithms should be based. Establishing a regulatory framework for algorithmic transparency and accountability is an *ex post* measure. Considering that algorithmic systems are "assemblage[s] of human and non-human actors" (Ananny and Crawford, 2018, p. 974) composed of code, practices, and norms, it is time to examine the norms or values algorithms should seek to reinforce. Regarding news algorithms, in particular, the following question – raised several years ago – remains relevant because scarcely any relevant research has yet been published: "Can we design and implement algorithms that optimize for an informed public, rather than page views and traffic?" (Lotan, 2014, p. 118). More specifically, can we develop news algorithms that rank articles with greater journalistic value higher?

This study provides initial responses to these questions. Current news algorithms, which have developed distinctly from journalism, remind us of the importance of journalistic values. We attempted to identify the ways journalistic values can be computationally quantified to make them readable by algorithms. As a first step to grapple with this question, we matched the textual indices (extracted from NLP/automated content analysis) with human conceptions of journalistic values (derived from survey analysis) by implementing PLS path modeling.

We found that the numbers of words or quotes have a strong association with the evaluations of balance, diversity, importance and factuality. This finding implies that to be regarded as high-quality journalism, news articles must be of a certain length and supply a certain number of citations supporting the story. In addition, we found that the linguistic polarization is an inverse indicator of respondents' perception of balance, diversity and importance – the more negatively or positively polarized news articles are, the lower the likelihood that respondents will evaluate them as balanced, diverse and important. Our analysis also shows that linguistic intensity is not a good indicator for respondents' perceptions of importance and factuality, though it is useful for gauging their perception of sensationalism – the less intense the words, the lower the likelihood that respondents assess news articles as sensational. The numbers of adverbs and adjectives were also useful

for estimating respondents' perceptions of factuality and sensationalism – respondents assess news articles as more sensational and less factual when they contain many flowery words. Our findings also indicate that news headlines play an important role in respondents' judgments of journalistic values. The presence of quotes, pair quotes, exclamation marks and question marks in news headlines are consistently associated with respondents' perception of lower journalistic values and increased sensationalism. All these findings are statistically significant even after controlling for news content type (soft/non-soft).

As expected, based on our review of previous studies of credibility and journalistic values, we found that perceived diversity, importance, factuality, readability and sensationalism determine respondents' news credibility perception. This indicates that respondents assess news articles that they believe fulfill journalistic values as credible – a well-accepted presumption (Hayes *et al.*, 2007) that had not been explicitly tested prior to this study. Interestingly, balance was not a significant predictor of respondent perceptions of credibility. The data set used in this study may not have included news topics that sufficiently triggered pro and con positions to elicit this effect. Alternatively, balance may no longer be important if news articles fairly treat diverse stakeholders or opinions, contain many factual cues, address important issues and deliver readable stories.

Regarding news types, we found non-soft news to be perceived as more balanced, diverse, important, factual and less sensational than soft news. Interestingly, our analysis showed that respondents perceive non-soft news as more readable than soft news. Since the issues that non-soft news stories address – politics, business and international affairs – are more complicated and require prior knowledge, we expected soft news to be more readable. This unexpected finding could result from the fact that non-soft news articles are written in a more logical and streamlined manner – a possibility the NLP technique and automated content analysis used in this study prevented us from testing. Alternatively, additional hidden textual features of which we were unaware may have influenced respondent perceptions of readability. As such, readability was the least explained concept in the present model, which highlights the need for additional research.

As an exploratory approach to a question that has not been previously investigated, this study naturally had some limitations. The survey respondents were graduate students majoring in journalism who were younger, highly educated and probably more news literate than general citizens, which prevents us from making any generalization beyond this group. In addition, the very limited number of survey respondents evaluating the journalistic values and credibility of news articles also makes it impossible for us to guarantee the representativeness of responses. Since this survey demanded considerable time and effort, we inevitably ended up having a low number of respondents. To address this limitation, we implemented several statistical treatments such as using *z*-scores and cutting off cases that revealed certain disagreements between respondents.

Despite these limitations, however, our findings may have implications for credibility research, news algorithm development and computational journalism research – the three realms with which this study intersects. The question of how journalistic values relate to perceptions of credibility requires additional scholarly attention, especially considering that few studies have analyzed the credibility of messages as opposed to the credibility of sources and mediums (Sundar, 1999). Our findings suggest a close link between journalistic value perception and news credibility assessment. This link needs further investigation by testing it against a larger sample of general citizens. Researchers have observed a divergence in the news choices of journalists and general citizens (Boczkowski and Mitchelstein, 2015). News algorithms that favor eye-catching and sensational news stories may have facilitated this gap. Future research should further investigate how people's perceptions of news credibility are influenced by journalistic values and thus explain the apparent news preference gap between journalists and lay people.

In addition, based on our findings, news algorithm developers should consider designing algorithms to select and prioritize news articles that have more words and quotes; fewer intense words, adjectives and adverbs; lower side-taking (either positive or negative) propensities; and fewer pair quotes, quotes, exclamation marks and question marks in the headlines. Consider, for example, a scenario in which we have two non-soft news articles. Both articles A and B have exactly the same number of words and headline quotes. Article A is less inclined to take either a positive or negative side and has more quotes than article B (i.e. A's textual indices generate a higher diversity score than B's), and article B has fewer adjectives and adverbs than A (i.e. B's textual indices generate a higher factuality score than A's). In this case, the news algorithm should be designed to value article A more than article B, because our findings show that diversity explains credibility more than factuality. Of course, our findings need to be substantiated with additional evidence.

Future studies should survey a much greater number of respondents to produce more generalizable results regarding people's perceptions of journalistic values and news credibility; going forward, researchers should also computationally identify a greater number of textual indices that represent humans' journalistic values. Knowledge of linguistics, NLP, text mining techniques, deep learning and other computational techniques may facilitate more sophisticated analyses that allow researchers to capture not only the words but also the contexts of texts. We did not use such techniques in this study because they operate like black boxes and are more suited for practical prediction than social-scientific explanation. Since the quantification of journalistic values and their association with news credibility has not been examined previously, we decided that, at this stage, traditional social science methods may produce more meaningful contributions. To advance the field of computational journalism, future studies should involve close collaborations with computer scientists and linguistics researchers.

A gap currently exists between the journalistic values that buttress our democratic decision-making processes and the operation of news algorithms developed by engineers and digital platform owners. The rise of artificial intelligence highlights the importance of examining how human values can be quantified and taught to computers; such examinations will help us better understand once-abstract human values and contribute to the creation of controllable and explainable computational models. Especially, journalism studies researchers should consider undertaking such examinations, since in the contemporary news environment news algorithms exert far more influence on general citizens' daily news consumption than journalistic news production – indeed, these days, journalists' news articles rarely reach online news users if news algorithms do not rank them at the top of search results. We hope this study reveals one possible direction social science can take in the era of the so-called fourth industrial revolution.

Notes

1. This study focuses on "news curation with algorithms," not on "news creation by algorithms." We focus on news distribution using news algorithms embedded in news aggregation sites, social media and other online news venues.
2. The combination of proprietary algorithms assessing relevance, opaque processes of human adjudication, and the lack of any visible public discussion leaves critical decisions about difficult content in the hands of a few unknown figures at social media companies (Crawford and Gillespie, 2016, p. 424).
3. This information was retrieved from the homepage of the American Society of Newspaper Editors on January 25, 2017: http://asne.org/content.asp?pl=24&sl=171&contentid=171

4. This information was retrieved from the homepage of the BBC on January 25, 2017: http://downloads.bbc.co.uk/aboutthebbc/insidethebbc/howwework/reports/pdf/neil_report.html
5. This information was retrieved from the homepage of the Canadian Broadcasting Corporation on January 25, 2017: www.cbc.radio-canada.ca/en/reporting-to-canadians/acts-and-policies/programming/journalism/
6. This information was retrieved from journalism.org on January 25, 2017: www.stateofthemedia.org/2005/newspapers-intro/content-analysis/
7. www.kpf.or.kr
8. All the text variables showed extremely positive skewness, ranging from 0.2 to 8.0 with a standard error of 0.08. Among the journalistic value variables, factuality, readability and credibility were negatively skewed (−0.6~−0.3), and the others were positively skewed (0.5~1.7).

References

Ananny, M. (2016), "Toward an ethics of algorithms convening, observation, probability, and timeliness", *Science, Technology & Human Values*, Vol. 41 No. 1, pp. 93-117.

Ananny, M. and Crawford, K. (2018), "Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability", *New Media & Society*, Vol. 20 No. 3, pp. 973-989.

Bakker, P. (2012), "Aggregation, content farms and Huffinization: the rise of low-pay and no-pay journalism", *Journalism Practice*, Vol. 6 No. 56, pp. 627-637.

Boczkowski, P.J. and Mitchelstein, E. (2015), *The News Gap: When the Information Preferences of the Media and the Public Diverge*, MIT Press, Cambridge, MA.

Bucher, T. (2017), "'Machines don't have instincts': articulating the computational in journalism", *New Media & Society*, Vol. 19 No. 6, pp. 918-933.

Cassidy, W.P. (2007), "Online news credibility: an examination of the perceptions of newspaper journalists", *Journal of Computer-Mediated Communication*, Vol. 12 No. 2, pp. 478-498.

Choi, S. and Kim, J. (2017), "Online news flow: temporal/spatial exploitation and credibility", *Journalism*, Vol. 18 No. 9, pp. 1184-1205.

Cohen-Almagor, R. (2008), "The limits of objective reporting", *Journal of Language and Politics*, Vol. 7 No. 1, pp. 136-155.

Crawford, K. and Gillespie, T. (2016), "What is a flag for? Social media reporting tools and the vocabulary of complaint", *New Media & Society*, Vol. 18 No. 3, pp. 410-428.

Davies, N. (2009), *Flat Earth News (An Award-winning Reporter Exposes Falsehood, Distortion and Propaganda in the Global Media)*, Vintage, Random House, London.

Deuze, M. (2005), "What is journalism? Professional identity and ideology of journalists reconsidered", *Journalism*, Vol. 6 No. 4, pp. 442-464.

Deuze, M. and Yeshua, D. (2001), "Online journalists face new ethical dilemmas: lessons from the Netherlands", *Journal of Mass Media Ethics*, Vol. 16 No. 4, pp. 273-292.

Diakopoulos, N. (2015), "Algorithmic accountability: journalistic investigation of computational power structures", *Digital Journalism*, Vol. 3 No. 3, pp. 398-415.

Diakopoulos, N. and Koliska, M. (2017), "Algorithmic transparency in the news media", *Digital Journalism*, Vol. 5 No. 7, pp. 809-828.

Dörr, K.N. and Hollnbuchner, K. (2017), "Ethical challenges of algorithmic journalism", *Digital Journalism*, Vol. 5 No. 4, pp. 404-419.

Garson, D. (2016), *Partial Least Squares: Regression and Structural Equation Models. Statistical Associates Blue Book Series*, Statistical Associates Publishing, Asheboro, NC.

Gillespie, T. (2014), "The relevance of algorithms", in Gillespie, T., Boczkowski, P.J. and Foot, K.A. (Eds), *Media Technologies: Essays on Communication, Materiality, and Society*, The MIT Press, Cambridge, MA, pp. 167-194.

Gitlin, T. (1980), *The Whole World is Watching: Mass Media in the Making & Unmaking of the New Left: With a New Preface*, University of California Press, Berkeley, CA.

Gladney, G.A. (1996), "How editors and readers rank and rate the importance of eighteen traditional standards of newspaper excellence", *Journalism & Mass Communication Quarterly*, Vol. 73 No. 2, pp. 319-331.

Gladney, G.A., Shapiro, I. and Castaldo, J. (2007), "Online editors rate web news quality criteria", *Newspaper Research Journal*, Vol. 28 No. 1, pp. 55-69.

Goffman, E. (1974), *Frame Analysis: An Essay on the Organization of Experience*, Harper & Row, New York, NY.

Hayes, A.S., Singer, J.B. and Ceppos, J. (2007), "Shifting roles, enduring values: the credible journalist in a digital age", *Journal of Mass Media Ethics*, Vol. 22 No. 4, pp. 262-279.

Holton, A.E., Coddington, M. and Gil de Zúñiga, H. (2013), "Whose news? Whose values? Citizen journalism and journalistic values through the lens of content creators and consumers", *Journalism Practice*, Vol. 7 No. 6, pp. 720-737.

Johnson, T.J. and Kaye, B.K. (2014), "Credibility of social network sites for political information among politically interested internet users", *Journal of Computer-Mediated Communication*, Vol. 19 No. 4, pp. 957-974.

Just, N. and Latzer, M. (2017), "Governance by algorithms: reality construction by algorithmic selection on the internet", *Media, Culture & Society*, Vol. 39 No. 2, pp. 238-258.

Karlsson, M. (2010), "Rituals of transparency: evaluating online news outlets' uses of transparency rituals in the United States, United Kingdom and Sweden", *Journalism Studies*, Vol. 11 No. 4, pp. 535-545.

Klinenberg, E. (2005), "Convergence: news production in a digital age", *The Annals of the American Academy of Political and Social Science*, Vol. 597 No. 1, pp. 48-64.

Kovach, B. and Rosenstiel, T. (2007), *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect*, Three Rivers Press, New York, NY.

Lanham, R.A. (2006), *The Economics of Attention: Style and Substance in the Age of Information*, University of Chicago Press, Chicago, IL.

Lotan, G. (2014), "Networked audiences: attention and data-informed", in McBride, K. and Rosenstiel, T. (Eds), *The New Ethics of Journalism: Principles for the 21st Century*, CQ Press, Sage, Thousand Oaks, CA, pp. 105-122.

Mitchell, A., Gottfried, J., Barthel, M. and Shearer, E. (2016), "Trust and accuracy", Pew Research Center, Washington, DC, available at: www.journalism.org/2016/07/07/trust-and-accuracy/ (accessed November 1, 2018).

Nielsen, J. (2006), "F-shaped pattern for reading web content", Nielsen Norm Group, available at: www.nngroup.com/articles/f-shaped-pattern-reading-web-content/ (accessed February 6, 2017).

Oh, S. (2016), "An exploratory inquiry into the convergence of journalism and algorithm", *Korean Journal of Cybercommunication Academic Society*, Vol. 33 No. 3, pp. 51-101.

Oh, S. and Kim, S. (2016), *Technological Proposal for Digital Journalism Transparency*, Korea Press Foundation, Seoul.

Ringle, C.M., Wende, S. and Becker, J.M. (2015), *SmartPLS 3*, SmartPLS, Boenningstedt.

Sallot, L.M., Steinfatt, T.M. and Salwen, M.B. (1998), "Journalists' and public relations practitioners' news values: perceptions and cross-perceptions", *Journalism & Mass Communication Quarterly*, Vol. 75 No. 2, pp. 366-377.

Sambrook, R.J. (2012), *Delivering Trust: Impartiality and Objectivity in the Digital Age*, Reuters Institute for the Study of Journalism, Oxford.

Segev, E. (2017), "From where does the world look flatter? A comparative analysis of foreign coverage in world news", *Journalism*, doi: 10.1177/1464884916688292.

Sherman, C. (2005), "A new F-word for Google search results", Search Engine Watch, available at: https://searchenginewatch.com/sew/news/2066806/a-new-f-word-google-search-results (accessed February 6, 2017).

Shin, H., Kim, M., Jo, Y., Jang, H. and Cattle, A. (2012), "Annotation scheme for constructing sentiment corpus in Korean", *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*, pp. 181-190.

Singer, J.B. (2010), "Norms and the network: journalistic ethics in a shared media space", in Meyers, C. (Ed.), *Journalism Ethics: A Philosophical Approach*, Oxford University Press, New York, NY, pp. 117-129.

Stark, J.A. and Diakopoulos, N. (2016), "Towards editorial transparency in computational journalism", *Computation + Journalism Symposium*, *September 30-October 1*, *Palo Alto, CA*.

Steele, B. (2008), "Ethical values and quality control in the digital era", *Nieman Reports*, Vol. 62 No. 4, pp. 57-58.

Sundar, S.S. (1999), "Exploring receivers' criteria for perception of print and online news", *Journalism & Mass Communication Quarterly*, Vol. 76 No. 2, pp. 373-386.

Tsang, E.W. (2002), "Acquiring knowledge by foreign partners from international joint ventures in a transition economy: learning-by-doing and learning myopia", *Strategic Management Journal*, Vol. 23 No. 9, pp. 835-854.

Van Der Wurff, R. and Schönbach, K. (2011), "Between profession and audience: codes of conduct and transparency as quality instruments for off-and online journalism", *Journalism Studies*, Vol. 12 No. 4, pp. 407-422.

Vinzi, V.E., Trinchera, L. and Amato, S. (2010), "PLS path modeling: from foundations to recent developments and open issues for model assessment and improvement", in Vinzi, V.E., Chin, W. W., Henseler, J. and Wang, H. (Eds), *Handbook of Partial Least Squares*, Springer, Berlin and Heidelberg, pp. 47-82.

Wilson, E.J. and Sherrell, D.L. (1993), "Source effects in communication and persuasion research: a meta-analysis of effect size", *Journal of the Academy of Marketing Science*, Vol. 21 No. 2, pp. 101-112.

Wong, K.K.K. (2013), "Partial least squares structural equation modeling (PLS-SEM) techniques using SmartPLS", *Marketing Bulletin*, Vol. 24 No. 1, pp. 1-32.

Corresponding author
Sujin Choi can be contacted at: sujinchoi2012@gmail.com

Social media analytics: analysis and visualisation of news diffusion using NodeXL

Wasim Ahmed
Northumbria University, Newcastle upon Tyne, UK, and
Sergej Lugovic
Zagreb University of Applied Sciences, Zagreb, Croatia

Received 15 March 2018
Revised 18 June 2018
15 September 2018
Accepted 17 September 2018

Abstract

Purpose – The purpose of this paper is to provide an overview of NodeXL in the context of news diffusion. Journalists often include a social media dimension in their stories but lack the tools to get digital photos of the virtual crowds about which they write. NodeXL is an easy to use tool for collecting, analysing, visualising and reporting on the patterns found in collections of connections in streams of social media. With a network map patterns emerge that highlight key people, groups, divisions and bridges, themes and related resources.
Design/methodology/approach – This study conducts a literature review of previous empirical work which has utilised NodeXL and highlights the potential of NodeXL to provide network insights of virtual crowds during emerging news events. It then develops a number of guidelines which can be utilised by news media teams to measure and map information diffusion during emerging news events.
Findings – One emergent software application known as NodeXL has allowed journalists to take "group photos" of the connections among a group of users on social media. It was found that a diverse range of disciplines utilise NodeXL in academic research. Furthermore, based on the features of NodeXL, a number of guidelines were developed which provide insight into how to measure and map emerging news events on Twitter.
Social implications – With a set of social media network images a journalist can cover a set of social media content streams and quickly grasp "situational awareness" of the shape of the crowd. Since social media popular support is often cited but not documented, NodeXL social media network maps can help journalists quickly document the social landscape utilising an innovative approach.
Originality/value – This is the first empirical study to review literature on NodeXL, and to provide insight into the value of network visualisations and analytics for the news media domain. Moreover, it is the first empirical study to develop guidelines that will act as a valuable resource for newsrooms looking to acquire insight into emerging news events from the stream of social media posts. In the era of fake news and automated accounts, i.e., bots the ability to highlight opinion leaders and ascertain their allegiances will be of importance in today's news climate.

Keywords Twitter, Social media, Social network analysis, Fake news, Information diffusion, Bots
Paper type Research paper

Introduction

Social media platforms generate substantial amounts of information on a range of topics and have become important channels of information flow in the twenty-first century (Bruns *et al.*, 2014). Originally, social media were created to allow members of the public to connect to one another for personal use but their use now stretches beyond this. One area that has benefited from the advent of social media is newsrooms because journalists are likely to include a social media element to their stories. Indeed, citizens may now also expect media sources to provide a social media element. One social media platform that has risen in popularity for its ability to diffuse news rapidly across the world is Twitter.

Twitter boasts an impressive 328m monthly active users with 1bn unique visits to tweets across the World Wide Web (About Twitter, n.d.), and Twitter is utilised by a subset of the global human population (Ahmed, 2017; Andrew Perrin, 2015; Holmberg and Thelwall, 2014). Due to the number of active users of Twitter emerging news events may appear to be reported on Twitter prior to reaching traditional media outlets. It is not surprising, therefore, that tweets in themselves have become to appear on television and

Online Information Review
Vol. 43 No. 1, 2019
pp. 149-160

1468-4527
DOI 10.1108/OIR-03-2018-0093

television news (Lefky *et al.*, 2015) and also within newspapers as direct sources of information in themselves (Brands *et al.*, 2018).

However, it can be argued that Twitter has been poorly mapped and understood for its network properties by news media. This is because although it is possible to visualise the structure of a conversation on Twitter and to identify prominent users and the overall structure of the conversation in order to garner the situational awareness of an emerging news story this aspect of Twitter is seldom reported on by news media. One software application which has gained in popularity for mapping and measuring content from social media platforms is known as NodeXL. NodeXL is a free, open source template for Microsoft Excel versions 2007, 2010, 2013 and 2016 which allows users to generate social network graphs. NodeXL Pro, a subscription, service, offers a number of additional features such as advanced network data streams, advanced network metrics, text and sentiment analysis, as well as powerful report generating. By making use of NodeXL it is possible to understand how communities form online, and to identify influential users as well as to pinpoint the content they share. NodeXL requires no specific technical knowledge and can be utilised by researchers from the social sciences.

NodeXL has been recently used to map and measure social media content during natural disasters in order to identify user interaction of communities (Daga, 2017). By doing so, insight can be gained into the interaction of Twitter users in an online community and can play a vital role in disseminating information during disaster and emergency situations. Similarly, news organisations can benefit by mapping online communities in order to gain insight into key information diffusers, popular sentiment and overall discussion. Previous research, for instance, has used NodeXL to map and measure the information diffusion of content surrounding emerging news stories such as the Occupy Wall Street movement (Park *et al.*, 2015).

The overall aim of this paper is to provide an overview of NodeXL and its potential for analysing and visualising news on Twitter. The objectives of the paper are to review the current uses of NodeXL in academic work including those from outside the area of social media, and then to specifically highlight the unique features of NodeXL for analysis and visualisation of news diffusion. The paper then proposes guidelines for newsrooms and journalist to grasp the situational awareness of emerging news events utilising NodeXL.

Literature review

The section will explore trends in this area which centre on the rise of news consumption through social media, challenges around data access, social media manipulation and the importance of having appropriate tools to analyse streams of social media data.

It can be argued that the importance of gaining insight into social media has been well established as early research has highlighted the global use of social media platforms and the potential to study inter-personal communication (Boyd and Ellison, 2007; Kaplan and Haenlein, 2010; Thelwall, 2009). In recent years, the influence of social media on society has been well studied (Fuchs, 2014) and in order to enable and encourage this form of research there have been calls for better tools for studying social media data (Ahmed, 2015). Scholars have noted the potential of such tools in the domain of social science research (Felt, 2016).

One issue that faces studying social media is that not all social media platforms make their data available to access for research purposes. Nine out of the top ten most used platforms significantly restrict data access (Ahmed, 2017) consequently making it a challenge to study them. There may be possibilities join "in-house" research teams at social media companies such as Facebook, but this is likely to be limited to a handful of researchers and can be an extremely competitive process.

One of the few social media platforms to provide near-complete public access to its data is Twitter and it can be argued that the infrastructure of Twitter is unique (Ahmed, 2018)

such that the majority of Twitter accounts are public (Marwick and Boyd, 2011). Furthermore, due to the ability to make use of hashtags on the platform that anyone can contribute to makes Twitter an ideal platform for emerging news stories. The "trending" feature on Twitter will display users topics attracting the most tweets regionally or nationally depending on the setting selected by a Twitter user (Kwak *et al.*, 2010). Newsrooms, therefore, are in a unique position to be able to rapidly analyse Twitter data during emerging events.

A recent trend to have emerged is the increased consumption of news through social media with 67 per cent of Americans indicating that they consume some form of news from social media (Shearer and Gottfried, 2017). One well-known case of news emerging through Twitter before traditional media outlets was the death of Osama Bin Laden which was leaked on the platform (Hu *et al.*, 2012). Moreover, Hu *et al.* (2012) noted that one of the reasons for Twitter users to become convinced of this was because the users who were posting the news appeared to be journalists and politicians, i.e., reputable individuals. Scholars have noted that users of social media platforms are likely to be influenced by opinion leaders whom play a critical role in how news is disseminated on social media (Bergström and Belfrage, 2018).

Twitter also has potential for citizen journalism because most smartphones are now able to capture an image on their device and have it uploaded to Twitter in under 45 s (Murthy, 2011). An iconic example of this is a passenger on the Midtown Ferry whom photographed a downed US Airways jet floating in the Hudson river in 2009 prior to the mainstream media even arriving to the scene (Murthy, 2011). These cases highlight the power of Twitter in the rapid cascading and diffusing information during emerging news events. Recent research has also found that untruthful information on social media has the potential to spread faster than truthful information (Vosoughi *et al.*, 2018).

Scholars have also found evidence of social media for "agenda setting", i.e., the ability to influence the news agenda of traditional media coverage (Feezell, 2018; Gidengil, 2014). This could potentially serve as a useful function as it may allow activists to raise awareness for legitimate causes. However, an area of concern in recent years is that social media manipulation is on the rise and it has been suggested that for certain topics related to Russia tweets which are produced by automated accounts could exceed 50 per cent (Stukal *et al.*, 2017). A recent study also found that Russian trolls on Twitter may promote discord around vaccines and share anti-vaccine content (Broniatowski *et al.*, 2018). A recent report by the Oxford Internet Institute from Oxford University has noted that computational propaganda is growing at a large rate (Bradshaw and Howard, 2018). The report noted that at least 48 countries have experienced some form of social media manipulation with half a billion dollars spent by political parties in areas such as psychological operations with the goal of manipulating public opinion (Bradshaw and Howard, 2018).

Therefore, with the rise of social media coupled with the occurrence of social media manipulation the ability to identify information diffusers and opinion leaders is ever more important for news organisations. However, to be able to critically study this content journalists will need access to tools that can analyse social media data. This paper will review literature around an emerging and exciting tool NodeXL which can pinpoint opinion leaders, and analyse overall content such as key topics, hashtags, websites and so forth. The paper will then outline how NodeXL can be applied by newsrooms by developing guidelines for its use.

Methods

In this study we utilised the Primo Central search engine at the University of Sheffield in order to locate literature, which includes over 17 journal databases including Scopus, MEDLINE, Springer Link, arXiv and Web of Science. In addition to using keyword search, citation

analysis of identified literature was also utilised (i.e. looking up references in bibliographies). Search terms were also entered into Google and Google Scholar. The date parameters for locating literature were from 2007 to 2018. The search terms utilised to locate literature on NodeXL's use in relation to social media consisted of "NodeXL", "NodeXL AND Twitter", "NodeXL AND Facebook", "NodeXL AND LinkedIn", and "NodeXL AND Instagram".

Terms that are more specific were used to identify potential different disciplines that may have used NodeXL and these terms consisted of "NodeXL AND Health", "NodeXL AND Politics", "NodeXL AND News", "NodeXL AND Disasters", "NodeXL AND Natural Disasters", and "NodeXL AND Scholarly Communication". Our inclusion criteria for studies were that:

- the study would use NodeXL to analyse data;
- the study was written in English;
- the paper could be accessed online; and
- the study was published in a journal and/or conference paper.

Furthermore to the aforementioned strategy above, two authors downloaded and reviewed 100 papers from the most popular citation from Google Scholar (Smith *et al.*, 2010) based on the inclusion criteria above. In total, 24 papers met the criteria and the topics that they analysed were extracted.

Results

By conducting a literature review it was found that there is a wide range of disciplines that have utilised NodeXL for the analysis of research data. NodeXL was most recently mentioned in an article published in *The Lancet*, a flagship medical journal, for its ability to analyse social media data related to infectious disease outbreaks (Mackenzie, 2018). The article noted that NodeXL reports were able to display popular content and provide insight into popular webpages that were cited in tweets. The use of NodeXL was also recently highlighted in a study which identified polarised crowds and opinion leaders in topics related to a viral hashtag on Twitter around abortion titled "#ShoutYourAbortion" (Ahmed, 2018). NodeXL in academic work has also ranged from the analysis of blackboard discussion boards (Waters and Gasson, 2012), blogs (Saffer, 2013), personal e-mail networks (Bengfort and Xirogiannopoulos, 2015), internal security (Gamachchi *et al.*, 2017), medical data analysis (Piepoli *et al.*, 2012), political sponsorship (Piepoli *et al.*, 2012), news stories (Quinn and Powers, 2016), campus-based social media platforms (Tang *et al.*, 2011) and Wikipedia (Ferron and Massa, 2011).

The majority of studies which have cited NodeXL would use the programme for the analysis of Twitter data, and these would form around a diverse range of fields and topics. A study (Tremayne and Minooie, 2013) utilised Twitter data using NodeXL in order to study opinion leadership around gun control shorty after gun-related violence in the USA. Others have studied political movements such as the Occupy Wall Street movement (Tremayne, 2014), health topics such as childhood obesity (Harris, Moreland-Russell, Tabak, Ruhr and Maier, 2014), UK local authorities (Panagiotopoulos and Sams, 2012), natural disasters, health (Harris, Moreland-Russell, Choucair, Mansour, Staub and Simmons, 2014), information diffusion (Russell *et al.*, 2015), trust on social networks (Faisal *et al.*, 2014), libraries (Borgatti *et al.*, 2014) and political language use (Piepoli *et al.*, 2012) have also all been studied.

An important aspect to social media for news diffusion is its ability to map the situational awareness of protest movements that typically receive a burst of social media attention. This is because key protesters will play a significant role during a protest by diffusing information and disseminating specific types of information. For example, a study

(Zamir, 2014) would use NodeXL in order to map discussion around the 2013 Shahbag movement in Bangladesh which refers to a series of protests following demands for the capital punishment of a number of individuals convicted of war crimes. Using NodeXL, it is possible to capture virtual crowds that may gather on social media platforms on Twitter via the use of network and graph theories. By doing so, the study found that those users who were influential during this time period consisted of citizens and journalists.

A further study would use NodeXL to evaluate information cascades that would take place in online discussion related to the #RaceTogether campaign (Feng, 2016). In using NodeXL it was possible to identify five different types of influential users such as those starting conversations, those influencing discussion, active engagers, network builders and users who were acting as a bridge for information. NodeXL is particularly apt for studying hashtags and a further recent study utilised NodeXL to study abortion content and found that Twitter discussion would form in two polarised groups such as users whom are either pro-abortion and those that are anti-abortion with low interaction between the groups (Ahmed, 2018).

NodeXL has also been applied in helping libraries promote their services, for instance, one study would use NodeXL to identify influential accounts connected to the libraries and would then strategically create and disseminate content for maximum exposure (Shulman *et al.*, 2015). The study by Shulman *et al.* (2015) also highlights how NodeXL could be utilised by news media organisations and journalists for identifying influential accounts and finding ways to share content with them which could potentially increase the reach of news articles.

General elections are likely to be of interest to news organisations as they are often reported on; and in one study NodeXL was utilised to study network structures during the German National Election campaign in 2017 (Reinhardt, 2018). Political topics more generally have also been studied and Choi *et al.* (2014), for instance, explored political discussions specifically for Korean Twitter users. NodeXL has also been utilised to study advocates and critics of the Hijab (a veil typically worn by women) on Twitter (Batet and Sánchez, 2018).

In an ever increasing and data-driven world where social media platforms are generating swarms of content it is important to have access to tools, methods and techniques, which can be used to make sense of this data. These findings from the review of literature highlight the diverse potential of social media data for gaining insight into a far-reaching range of topics which could be applied to the news media domain.

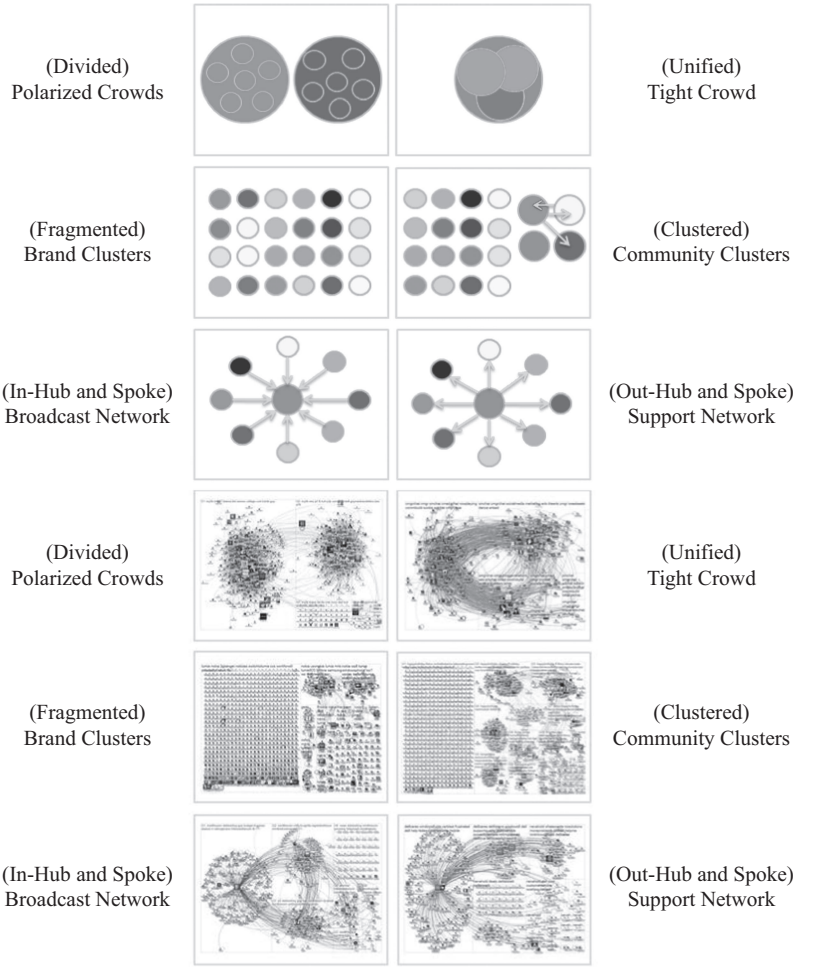
It is also important to provide an outline the functional capability of NodeXL and some of the features it contains. NodeXL allows end-users to generate network visualisations from a range of data sources and one such source is Twitter. In the case of Twitter NodeXL can additionally generate a number of metrics associated with the graphs such as:

- The most frequently shared URLs.
- Domains.
- Hashtags.
- Words.
- Word Pairs.
- Replied-To.
- Mentioned Users.
- Most frequent tweeters.

These metrics are produced overall and also by group of Twitter users. This is because NodeXL has the ability to cluster discussion into a number of different groups based on the content that is shared. Thus, looking at different metrics associated with different groups (G1, G2, G3, etc.), it is possible to rapidly establish the diverse areas that users may be

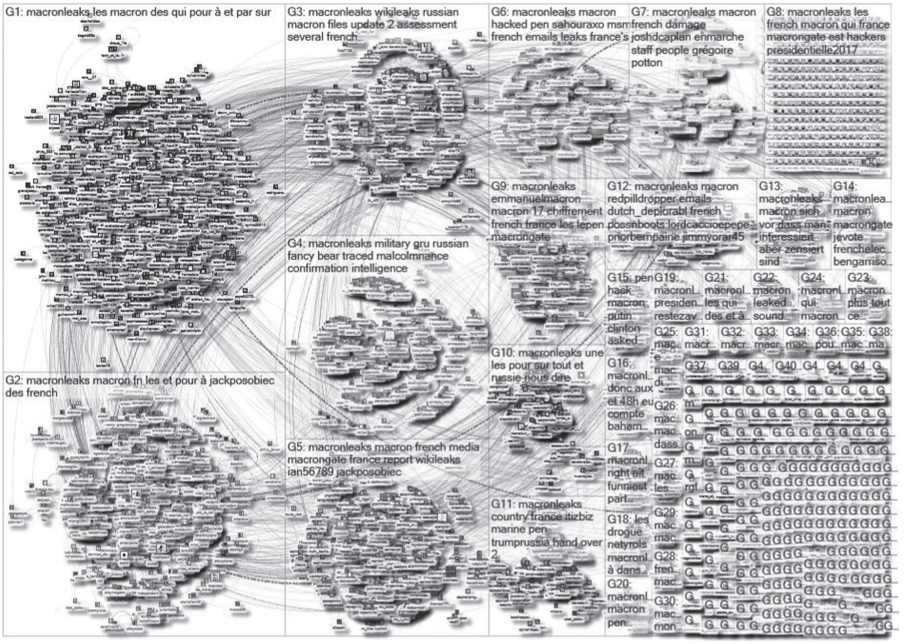
conversating about. NodeXL also hosts an online "Graph Gallery" where users can upload workbooks and network graphs. Figure 1 showcases the different network structures that can emerge on Twitter, and Figure 2 provides a simplified view on the visualisation.

Figure 1, from Smith *et al.* (2014), highlights how distinct topics on social media social media can have contrasting network patterns. For instance in the polarised crowd discussion this may occur when the topics are split into two sets of users referring to two distinct topics, for instance, one set of users may converse about Donald Trump and another set of users may form their own group to converse about Hilary Clinton. The polarised crowd groups can be identified because two distinct groups can be observed within a network graph. Further examples could include Twitter users who are pro-abortion and users whom are anti-abortion when discussion surrounding abortion is analysed. In the unified crowd, referring back the political analogy, users may converse about different aspects of an election which contains some overlap as certain users may be connected to each other. A unified crowd network is typical at academic conferences because discussion



Source: Smith *et al*. (2014)

Figure 1.
Network structures that emerge on Twitter



Note: Network analysis of #Macronleaks using NodeXL

Figure 2.
Network structures that emerge on Twitter (simplified view)

may overlap and Twitter users may already be connected to each other. The brand clusters shape would occur when users would, for instance, share an opinion or website link without being connected to one another. Popular news events which generate news articles may contain a sizeable isolates group because Twitter users will repost these articles. In community clusters users may form smaller groups to discuss sub-topics related to an overall topic. For instance, given discussion on a sports event Twitter users may begin new sub-topics related to the specific performance of certain players. Broadcast networks are typically found when analysing news accounts as these disseminate news which is retweeted by a large amount of users. In emerging news events newsrooms may be interested to examine the key users tweeting about a topic in order to determine key information diffusers. These users may be present among several subgroups and may contain news accounts, journalists and opinion leaders. In the UK, the Twitter networks of BBC News and Sky News, for instance, will resemble that of a broadcast network because their tweets are widely shared and retweeted. Support networks are those accounts which reply to a large number of accounts, for instance, the customer support of a bank may reply to a large amount of Twitter users to address their queries.

Reporting network metrics and visualisations

Table I connects potential goals of a newsroom to specific features in NodeXL which can be utilised to garner insight into discussion on Twitter. This is the first empirical study which explicitly highlights and connects overarching goals of newsrooms to specific outputs and our guidance is likely to have important practical uses during emerging events for those who work in the news domains.

The goals listed above were utilised alongside NodeXL to study a leak of e-mails during the 2017 French presidential elections, the narratives that were shared, and the users that

General goals for newsrooms	How to achieve goal	Example
Determine dominant external media narratives shared on social media during an evolving news event. And establish different discussions that are taking place based on an emergent new development	Examine most frequently shared URLs, domains and hashtags in NodeXL. And examine the different groups by examining and interpreting the most frequently occurring words, word pairs in order to understand the discussions that are taking place	During a presidential election one of the candidates has their e-mails leaked. WikiLeaks (an international non-profit organisation which shares private information) shares this content on social media. News media may be interested in finding out the most shared URLs linking to the leaked e-mails, the domains referred to, and the hashtags that may be referred to
Ascertain users who are concerned with an evolving news event	Examine influential Twitter users by the metric of Betweenness Centrality, Indegree and OutDegree	In a similar example as above, news media organisations may wish to identify users on social media platforms such as Twitter who hold influence. That is users who are influential can be measured by Betweeness Centrality, users tweeting the most will have a high OutDegree and users being mentioned the most will have a high Indegree

Table I.
Goals for news teams analysing Twitter data

were influential (Ahmed and Downing, 2017). Figure 2 provides an overview of the network visualisation that was created by analysing discussion on this topic.

In this specific study by mapping the network of the #Macronleaks event (as shown in Figure 2) it was possible to identify Twitter users who were most influential. The most influential user consisted of the Twitter account by Wikileaks. A further Twitter user who appeared to be most influential was a Donald Trump supporter (as indicated from their Twitter bio) from the USA among a number of alt right users. The visualisation indicates how Twitter users in different groups have specific conversations in which they converse about, for instance, the keyword "Russian" appears among the most used words because users would discuss whether the e-mail hack was orchestrated by Russian hackers. Furthermore, the study found that a top word pair combination, i.e., two words appearing together the most was "WikiLeaks" and "Macronleaks" which indicated the importance of Wikileaks as an information source which operates outside of the bounds of traditional journalism. The most frequently used URLs relate to an independent blog covering the news story because at the time there was a media blackout preventing mainstream media reporting on the story. The desktop version of NodeXL will display a list of "top items" such as hashtags, words, websites and influential users among other top metrics. These items combined with the visualisation can also be uploaded to the NodeXL graph gallery which allows users to freely share their data and outputs and will be a useful resource for those working in the news media domain.

The earlier sections of this paper highlighted the pressing concerns surrounding social media manipulation and the power of opinion leaders in disseminating news. Table I outlines how it would be possible to determine dominant external media narratives, key users and resources which would have the potential to highlight any potential social media manipulation.

Discussion

Social media has significantly altered the way in which citizens consume news and it has had a profound impact within the domains of social science research. Originally intended for

personal usage social media began to be utilised in news and academic research. Indeed, there has been an exponential burst in the way social media has begun to be utilised as a means of communication. Moreover, its swiftness in reporting breaking news stories has significantly altered the way in which journalism is conducted. It has also provided citizens with the power to become creators of news and has increased the speed in which news is consumed. These developments have provided an ability to gather information and data on various emerging news topics. However, this paper argued that news on social media has been poorly mapped and measured and provided an overview and guidelines for measuring and mapping collections of connections from social media platforms. NodeXL utilised alongside these guidelines could be put to use by journalists and researchers alike in developing a deeper understanding of social media reaction. This paper was the first empirical study which reviewed literature on NodeXL, and provided insight into the vast value of network visualisations as well as analytics for the news media domain. In the twenty-first century the way citizens communicate has significantly altered and new forms of technology such as social media has become common place. Henceforth, by examining content on social media dominant narratives can emerge and the users who hold the most power in regards to shaping the content of discussion can be understood. In an ever changing political landscape where social media is becoming ever more important in shaping political opinions and indeed potentially influencing the public it is vitally important to map and measure media narratives and influencers. Thus, this paper is likely to serve as a valuable resource for newsrooms whom can utilise the methods and techniques to better understand online content. By doing so news providers can also include a social media element to their stories and enrich content by providing insight into how online communities converse about certain topics and highlight how networks are structured. It must be noted that there are also further tools that can be used for social network analysis. Gephi (Bastian *et al.*, 2009), for instance, is an open source software, which can be utilised for the purposes of graph and network analysis. UCINET is a general package for social network analysis (Borgatti *et al.*, 2014). One of the benefits of using NodeXL is that it has a number of features which can automate the generation of the network visualisation and associated metrics. It is also worth mentioning a tool known as DiscoverText, a text-analytics tool, which can also provide key insight into social media data (Shulman, 2011).

Conclusion

This paper has provided an overview of some of the diverse uses of NodeXL from across a number of academic disciplines. The abstract provides an interesting starting point in reviewing the literature utilising NodeXL for research purposes. Now, as social media platforms become more and more popular it is critically important to study them. An interesting feature of NodeXL is that it requires no technical and/or programming knowledge which potentially allows it to be utilised across a wide range of disciplines from science and engineering, the social sciences and the humanities.

References

About Twitter (n.d.), "Twitter Q1 2017 Company Metrics", available at: https://about.twitter.com/company (accessed 19 June 2017).

Ahmed, W. (2015), "Using Twitter as a data source: an overview of current social media research tools", London School of Economics and Political Science Blog 10, London.

Ahmed, W. (2017), "Using Twitter as a data source: an overview of social media research tools (updated for 2017)", Impact of Social Sciences Blog.

Ahmed, W. (2018), "Public health implications of# ShoutYourAbortion", *Public Health*, Vol. 163, pp. 35-41.

Ahmed, W. and Downing, J. (2017), "Campaign leaks and the far-right: who influenced #Macronleaks on Twitter?", LSE European Politics and Policy EUROPP Blog, available at: http://blogs.lse.ac.uk/europpblog/ (accessed 16 June 2018).

Bastian, M., Heymann, S. and Jacomy, M. (2009), "Gephi: an open source software for exploring and manipulating networks", *International Conference on Web and Social Media*, Vol. 8, pp. 361-362.

Batet, M. and Sánchez, D. (2018), "Semantic disclosure control: semantics meets data privacy", *Online Information Review*, Vol. 42 No. 3, pp. 290-303.

Bengfort, B. and Xirogiannopoulos, K. (2015), *Visual Discovery of Communication Patterns in Email Networks*, Unpublished manuscript.

Bergström, A. and Belfrage, M.J. (2018), "News in social media: incidental consumption and the role of opinion leaders", *Digit Journal*, Vol. 6 No. 5, pp. 583-598.

Borgatti, S.P., Everett, M.G. and Freeman, L.C. (2014), "UCINET", in Alhajj, R. and Rokne, J. (Eds), *Encyclopedia of Social Network Analysis and Mining*, Springer, New York, NY.

Boyd, D.M. and Ellison, N.B. (2007), "Social network sites: definition, history, and scholarship", *Journal of Computer-Mediated Communication*, Vol. 13 No. 1, pp. 210-230.

Bradshaw, S. and Howard, P.N. (2018), "Challenging truth and trust: a global inventory of organized social media manipulation", Working Paper No. 1, Project on Computational Propaganda, Oxford Internet Institute, Oxford University, Oxford.

Brands, B.J., Graham, T. and Broersma, M. (2018), "Social media sourcing practices: how Dutch newspapers use tweets in political news coverage", *Managing Democracy in the Digital Age*, Springer, Cham, pp. 159-178.

Broniatowski, D.A., Jamison, A.M., Qi, S., AlKulaib, L., Chen, T., Benton, A. and Dredze, M. (2018), "Weaponized health communication: Twitter bots and Russian Trolls amplify the vaccine debate", *American Journal of Public Health*, Vol. 108 No. 10, pp. 1378-1384.

Bruns, A., Burgess, J.E., Mahrt, M., Puschmann, C. and Weller, K. (Eds) (2014), *Twitter and Society*, Peter Lang, New York, NY.

Choi, M., Sang, Y. and Woo Park, H. (2014), "Exploring political discussions by Korean Twitter users: a look at opinion leadership and homophily phenomenon", *Aslib Journal of Information Management*, Vol. 66 No. 6, pp. 582-602.

Daga, R.R.M. (2017), "Social network analysis of Tweets on Typhoon during Haiyan and Hagupit", *Proceedings of the 8th International Conference on Computer Modeling and Simulation, ACM*, pp. 151-154.

Faisal, M., Alsumait, A. and Zainab, A.A. (2014), "Trust inference algorithms for social networks", *Journal of Engineering Research*, Vol. 2 No. 2.

Feezell, J.T. (2018), "Agenda setting through social media: the importance of incidental news exposure and social filtering in the digital era", *Political Research Quarterly*, Vol. 71 No. 2, pp. 482-494.

Felt, M. (2016), "Social media and the social sciences: how researchers employ Big data analytics", *Big Data and Society*, Vol. 3 No. 1, p. 2053951716645828.

Feng, Y. (2016), "Are you connected? Evaluating information cascades in online discussion about the# RaceTogether campaign", *Computers in Human Behavior*, Vol. 54, pp. 43-53.

Ferron, M. and Massa, P. (2011), "Collective memory building in Wikipedia: the case of North African uprisings", *Proceedings of the 7th International Symposium on Wikis and Open Collaboration, ACM*, pp. 114-123.

Fuchs, C. (2014), *Social Media: A Critical Introduction*, Sage, London.

Gamachchi, A., Sun, L. and Boztas, S. (2018), A graph based framework for malicious insider threat detection.

Gidengil, E. (2014), *Canadian Democracy from the Ground up: Perceptions and Performance*, UBC Press, Vancouver.

Harris, J.K., Moreland-Russell, S., Tabak, R.G., Ruhr, L.R. and Maier, R.C. (2014), "Communication about childhood obesity on Twitter", *American Journal of Public Health*, Vol. 104 No. 7, pp. e62-e69.

Harris, J.K., Moreland-Russell, S., Choucair, B., Mansour, R., Staub, M. and Simmons, K. (2014), "Tweeting for and against public health policy: response to the Chicago department of public health's electronic cigarette Twitter campaign", *Journal of Medical Internet Research*, Vol. 16 No. 10.

Holmberg, K. and Thelwall, M. (2014), "Disciplinary differences in Twitter scholarly communication", *Scientometrics*, Vol. 101 No. 2, pp. 1027-1042.

Hu, M., Liu, S., Wei, F., Wu, Y., Stasko, J. and Ma, K.L. (2012), "Breaking news on Twitter", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, May*, pp. 2751-2754.

Kaplan, A.M. and Haenlein, M. (2010), "Users of the world, unite! The challenges and opportunities of social media", *Business Horizons*, Vol. 53 No. 1, pp. 59-68.

Kwak, H., Lee, C., Park, H. and Moon, S. (2010), "What is Twitter, a social network or a news media?", *Proceedings of the 19th International Conference on World Wide Web, ACM*, pp. 591-600.

Lefky, T., Brewer, P.R. and Habegger, M. (2015), "Tweets on television news: the nature and effects of Campaign coverage of Twitter", *Electronic News*, Vol. 9 No. 4, pp. 257-269.

Mackenzie, G. (2018), "Twitter big data and infectious disease conferences", *The Lancet Infectious Diseases*, Vol. 18 No. 2, p. 154.

Marwick, A. and Boyd, D. (2011), "To see and be seen: celebrity practice on Twitter", *Convergence*, Vol. 17 No. 2, pp. 139-158.

Murthy, D. (2011), "Twitter: microphone for the masses?", *Media, Culture & Society*, Vol. 33 No. 5, pp. 779-789.

Panagiotopoulos, P. and Sams, S. (2012), "An overview study of Twitter in the UK local government", *tGov*.

Park, S.J., Lim, Y.S. and Park, H.W. (2015), "Comparing Twitter and YouTube networks in information diffusion: the case of the 'Occupy Wall Street' movement", *Technological Forecasting and Social Change*, Vol. 95, pp. 208-217.

Perrin, A. (2015), "Social media usage", *Pew Research Center*, pp. 52-68.

Piepoli, A., Tavano, F., Copetti, M., Mazza, T., Palumbo, O., Panza, A. and Gentile, G. (2012), "Mirna expression profiles identify drivers in colorectal and pancreatic cancers", *PloS One*, Vol. 7 No. 3, p. e33663.

Quinn, K. and Powers, R.M. (2016), "Revisiting the concept of 'sharing' for digital spaces: an analysis of reader comments to online news", *Information, Communication and Society*, Vol. 19 No. 4, pp. 442-460.

Reinhardt, S. (2018), "Network Gatekeeping on Twitter during the German National Election Campaign 2017", Extended Abstract für das 14. Düsseldorfer Forum Politische Kommunikation vom 5-7 April.

Russell, S., Middleton-Green, L. and Johnston, B. (2015), "Using social media to create discussion", *International Journal of Palliative Nursing*, Vol. 21 No. 11, pp. 525-526.

Saffer, A.J. (2013), *Intermedia Agenda Building of the Blogosphere: Public Relations Role in the Network*, Manuscript submitted for publication.

Shearer, E. and Gottfried, J. (2017), "News use across social media platforms 2017", *Pew Research Center, Journalism and Media*.

Shulman, J., Yep, J. and Tomé, D. (2015), "Leveraging the power of a Twitter network for library promotion", *The Journal of Academic Librarianship*, Vol. 41 No. 2, pp. 178-185.

Shulman, S. (2011), "DiscoverText: software training to unlock the power of text", *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times, ACM*, p. 373.

Smith, M.A., Rainie, L., Shneiderman, B. and Himelboim, I. (2014), "Mapping Twitter topic networks: from polarized crowds to community clusters", *Pew Research Center*, Vol. 20, pp. 1-56.

Smith, M., Ceni, A., Milic-Frayling, N., Shneiderman, B., Mendes Rodrigues, E., Leskovec, J. and Dunne, C. (2010), "NodeXL: a free and open network overview, discovery and exploration add-in for Excel 2007".

Stukal, D., Sanovich, S., Bonneau, R. and Tucker, J.A. (2017), "Detecting bots on Russian political Twitter", *Big Data*, Vol. 5 No. 4, pp. 310-324.

Tang, T., Hämäläinen, M., Virolainen, A. and Makkonen, J. (2011), "Understanding user behavior in a local social media platform by social network analysis", *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, ACM*, pp. 183-188.

Thelwall, M. (2009), "MySpace comments", *Online Information Review*, Vol. 33 No. 1, pp. 58-76.

Tremayne, M. (2014), "Anatomy of protest in the digital era: a network analysis of Twitter and Occupy Wall Street", *Social Movement Studies*, Vol. 13 No. 1, pp. 110-126.

Tremayne, M. and Minooie, M. (2013), "Opinion leadership on gun control in social networks: preferential attachment versus reciprocal linking", *American Communication Journal*, Vol. 15 No. 4.

Vosoughi, S., Roy, D. and Aral, S. (2018), "The spread of true and false news online", *Science*, Vol. 359 No. 6380, pp. 1146-1151.

Waters, J. and Gasson, S. (2012), "Using asynchronous discussion boards to teach is: reflections from practice", *33rd International Conference on Information Systems*.

Zamir, M.H. (2014), "Diffusion of protest information in twitter during Shahbag Movement of Bangladesh", *Proceedings of the American Society for Information Science and Technology*, Vol. 51 No. 1, pp. 1-4.

Corresponding author
Wasim Ahmed can be contacted at: Wasim.Ahmed@Northumbria.ac.uk

The moderating effect of technology optimism

How it affects students' weblog learning

Cheng-Min Chao
Department of Business Administration, National Taichung University of Science and Technology, Taichung, Taiwan, and
Tai-Kuei Yu
Department of Business Administration, National Quemoy University, Kinmen Hsien, Taiwan

Received 4 November 2016
Revised 16 April 2017
28 May 2018
Accepted 15 July 2018

Abstract

Purpose – The purpose of this paper is to combine task-technology fit, theory of planned behaviour and individual technology optimism, and propose a better hybrid technology behavioural intention model to explain Taiwanese students' usage behaviour for weblog learning.

Design/methodology/approach – A 31-item questionnaire with eight constructs was administered to undergraduate and graduate students enrolled in three universities in Taiwan. A total of 380 voluntary, usable responses were received, and a research model estimated using Smart PLS was used to interpret the structural relation results.

Findings – The results of the research model were analysed using a structural equation modelling approach to test six hypotheses and three moderating hypotheses; significant support were found for seven of them. Accordingly, this study considered the level of technology optimism as a moderator to explore whether it impacts perceived behavioural control, attitudes and social influences on behavioural intention related to weblog learning.

Originality/value – This research provides a better understanding of individual and system characteristics, as well as social factors regarding weblog learning system acceptance and intention.

Keywords Theory of planned behaviour, Task-technology fit, Moderator effect, Technology optimism

Paper type Research paper

1. Introduction

Over the last 15 years, internet use has become more advanced, and there has been rapid growth and awareness of Web 2.0 technologies including blogs, microblogs and wikis. Such tools allow people to share opinions, thoughts and experiences with other users (Chen *et al.*, 2015; Chwo, 2015; Kang *et al.*, 2011; Mohammadyari and Singh, 2015). Many educators and researchers (Wang *et al.*, 2008; Yu *et al.*, 2010) claim that if students are constantly searching for knowledge, they will gradually better understand what they find. Applications such as: Wikipedia, Google, and Yahoo! Answers, allow internet users to discover knowledge through different ways. Meanwhile, learning systems such as weblogs, microblogs and wikis (Lee and Bonk, 2016; Chen *et al.*, 2015; Kang *et al.*, 2011; Wang *et al.*, 2008; Yu *et al.*, 2010) also help learners obtain knowledge and cultivate problem-solving skills. Among these applications, weblogs (often shortened to blogs) have become very prevalent, and are often used by experts and academics to express their perspectives and interact with others.

The use rate of weblogs is still expanding, and the user population has become more diverse (Chen *et al.*, 2015; Lu and Hsiao, 2009; Hsu and Lin, 2008; Thelwall and Hasler, 2007). Moreover, teaching or learning with blogs has become more popular over the last decade. The educational affordance of blogs and proposed frameworks for using blogs to learn have

The authors thank the National Science Council of Taiwan for financially supporting this research under contract MOST 100-2511-S-507-001-MY3.

Online Information Review
Vol. 43 No. 1, 2019
pp. 161-180

1468-4527
DOI 10.1108/OIR-11-2016-0316

been emphasised in several research studies (e.g. Wang *et al.*, 2016; Deng and Yuen, 2011; Robertson, 2011; Kim, 2008; Kerawalla *et al.*, 2009). Many researchers have investigated educational weblog behavioural aspects connected to learners' behaviour intentions, including theirs perceptions and attitudes about learning with blogs (e.g. Ladyshewsky and Gardner, 2008; Jimoyiannis and Angelaina, 2012). The aforementioned studies all focussed on depicting how blogs can be used in teaching and learning.

Weblogs are an online tool that enables users to communicate and share information (Lee and Bonk, 2016; Furukawa *et al.*, 2006). They are mainly used to support the acquisition and retrieval of codified knowledge in order to improve individual knowledge bases. Weblogs may be viewed as an evolved form of personal web pages used to publish personal knowledge. In the context of higher education, especially in business education, blogs are becoming more widely used, which can also potentially improve business students' communication skills, learning abilities and performance (Wang *et al.*, 2016; Freeman and Brett, 2012). Further, an increasing number of experts have started using blogs to present their work, follow developments in the field, and share their ideas, and thereby serve as an extension of the learning setting. In their study focussed on weblog learning in the fields of business and management education, Lee and Bonk (2016) proposed that blog owners may be motivated to be responsible for their learning with blogs due to the ownership factor. An important part of business and management education, e.g., marketing management, innovation management, entrepreneurial management and organisational behaviour is to inspire students' creative and interactive thinking, rather than subject them to one-way instruction. Hence, this study classifies weblog learning as an educational tool. With blogs, students can both enhance their personal learning and share their learning experiences and opinions with instructors and peers. Moreover, blogs allow students to post contents about assignments and discuss their thoughts on course materials.

In the literature about technology acceptance and adoption, a considerable number of models have been applied (e.g. the theory of reasoned action (TRA), the theory of planned behaviour (TPB) and the technology acceptance model) to investigate and explore the determinants of users' behaviour towards using information technology (IT) or systems. Among these models, the TPB is the most frequently cited and influential model in predicting and explaining one's intention/behaviour (Ajzen, 1991, 2011). This theory is an extension of the TRA, which considers both volitional and non-volitional aspects (Ajzen, 1991). Since its development, TPB has been extensively used, tested and extended to explain various human behaviours and successes in a number of application areas such as: online learning, mobile healthcare and information system (IS) (Ajzen, 1991; Wu *et al.*, 2011; Yu and Yu, 2010).

While such rational-choice theories as the TPB offer limited explanation of purposive intention/behaviour, most studies generally agree that these theories can be supplemented with several vital predictors (Yu and Yu, 2010). User acceptance and usage behaviour towards technology can be influenced by a variety of factors such as individual differences and social influence. However, weblog learning does not mean learning with new technology, but rather voluntarily using online platforms to gain knowledge and interact with other users around the globe. The TPB is applied to discuss the behaviour of technology adoption by considering the individual role and organisation systems in this course (Wu *et al.*, 2011). As the TPB does not explain variables, its ability to predict weblog learning might be limited; yet this limitation can be overcome by extending these two models with the task-technology fit (TTF) model, which is a commonly adopted theoretical model for assessment of how IT causes performance and usage impacts. For an IS to significantly affect technology application, the technology must fit the task it supports. Because the TTF model does not address social factors, it may have restricted ability to

predict weblog learning. Therefore, the TPB can be adopted to support the TTF. According to previous studies (Cheng *et al.*, 2012; Dishaw and Strong, 1999; Lee and Lehto, 2013; Närman *et al.*, 2012; Taylor and Todd, 1995; Yen *et al.*, 2010), the TPB and TTF have led to many positive results associated with investigations of IT usage and how it helps to improve work performance.

Over the past decade, in addressing the issue of users' adoption (utilisation) of new IS/IT, TTF models is important theoretical bases in the IS field (Yen *et al.*, 2010). The TTF proposed by Goodhue and Thompson (1995) stresses that one's performance, if well utilised, can be positively impacted by technology characteristics and task characteristics. Therefore, most previous studies have combined other theories based on the TTF model to study usage intention, user adoption (Lam *et al.*, 2007) and user behaviours associated with IT. Although the TTF model explains how technology can influence task performance (Goodhue and Thompson, 1995), most previous research on the TTF emphasised task and technology fit – the relationships among individuals, tasks and technology were often overlooked. In fact, for causes and effects of the technology-to-performance chain, it is notable that personal characteristics, along with task characteristics and technology characteristics are the antecedent variables of TTF (Goodhue and Thompson, 1995). Recently, many researchers have included personal factors as antecedent variables into the TTF model to further expand its interpretation of technology usage. However, these studies mostly focussed on the impact of personal factors on TTF (Jarupathirun and Zahedi, 2007; Teo, 2010). Little research has investigated both personal factors and TTF to study their relationship with personal work performance, as well as the relative impacts.

Since it was first proposed, the TTF model has been used to analyse various IS; several studies have addressed learners' intention to use system from the perspective of TTF (e.g. Goodhue, 1995; Goodhue and Thompson, 1995; Larsen *et al.*, 2009; Lee and Lehto, 2013; Lin, 2012; Ma *et al.*, 2013; Yen *et al.*, 2010; Yu and Yu, 2010; Wang *et al.*, 2016). However, little is known about how to evaluate educational IS, such as weblog learning sites, using the TTF model. Additional studies are needed to obtain more insights into its validation across different contexts and to determine whether a good TTF will affect user intentions of weblog learning, and the significance of this influence. The TTF model does not address social factors and may have a weaker ability in predicting weblog learning; therefore, the TPB can act as an extension.

In the past, e-learning was a one-sided learning method. Weblog learning solves this problem by allowing interactions in learning. Prior weblog research has empirically examined the effects of a large number of determinants on behavioural intention to use weblog learning systems, utilising a variety of theoretical models to explain individuals' behavioural intention to use the weblog learning system (Hsu and Lin, 2008; Lee and Lehto, 2013; Ma *et al.*, 2013; Mohammadyari and Singh, 2015; Yu and Yu, 2010; Wang *et al.*, 2016), e.g., cognitive aspects of technology acceptance, technology and task characteristics and IS success models, among others. The emergence of these theoretically derived approaches has also led to a large number of alternative models and extensions. Despite these results, few studies have investigated how one's readiness to use new technologies impacts students' usage behavioural intention, especially in the context of business and management education. In connection with the vitality of new learning technology or systems, users' attitudes, i.e., whether they accept new learning technologies or systems, is of growing interest to business and management educators who use new learning technologies and systems. Therefore, various studies attempted to understand users' technology readiness (TR) and effectively predict their behavioural intention (Parasuraman, 2000). To date, TR studies have been largely confined to the following domains: business to consumer (Vize *et al.*, 2013), social networking site (SNS) (Borrero *et al.*, 2014) and augmented reality (Chung *et al.*, 2015); few researchers have considered TR in the weblog learning systems context. Parasuraman (2000)

also emphasised the use of four personality traits to measure people's readiness to use new technologies: optimism and innovativeness (contributors) and discomfort and insecurity (inhibitors). The former two are enablers of new technology use: optimism refers to a positive attitude towards technology and a belief that it offers increased management control, flexibility and efficiency in one's life, while innovativeness refers to the tendency to be a technology pioneer and thought leader (Parasuraman, 2000). This study intends to understand students' feelings towards weblog learning and their perceptions regarding improved study efficiency and convenience. Therefore, this study considers technology optimism to have a relatively significant influence on students' usage of weblog learning. Whether technology optimism can moderate students' beliefs and behavioural intention is the centre of the study.

To summarise the above discussion, while the adoption process of weblog learning systems involves both technological and organisational aspects at the individual level, in this study, we integrate the TPB, technology optimism and the TTF in a complementary manner as the theoretical basis for the development of a new model towards weblog learning intention among Taiwanese students. Students learn about systems such as C++, java and VB, among others, and discuss related topics with other learners from all over the world. Additionally, this study argues that the relationship between students' expectancy beliefs about weblog learning and their intentional and actual use of them to expressively participate can be influenced by their technology optimism. For example, David is an information management student with high technology optimism who understands IT applications quite well, and is aware of the advantages of weblog learning over traditional classroom learning. In programming language courses, David uses a weblog as a complementary learning channel and conducts two-way discussions with other participants through it. As David's attitude towards weblog learning is more positive and his technology optimism is high, his learning progress accelerates. In contrast, the reverse of David's technology optimism will result in a rapid decline of his behavioural intention towards weblog learning. On the other hand, if David believes that weblog learning consists of numerous supporting technological and learning resources, he becomes even more likely to adopt weblog learning; combined with his high degree of technological optimism, this will further enhance his ability to adopt weblog learning and the relevant resources, as well as to reduce obstacles. Therefore, under the influence of the interactions of these variables, his intensity of perceived behavioural control will quickly increase. In addition, as David increasingly uses weblog learning, he often finds some weblog content that is either outdated or posted anonymously, raising concerns that the content is out of date or possibly not credible, as it is difficult to know if the posted contents are written by specialists or not. Even if David is a person with high technology optimism, considering the potential interactions of these variables, his weblog learning intentions are not affected by social influences of important persons.

This study adopts the TPB model (i.e. perceived behavioural control (PBC), attitude towards, social influences and behavioural intention) and adds external variables (i.e. individual characteristics, technology characteristics and system quality) to enhance the TPB model's ability to predict the determinants that affect students; behavioural intentions to use weblog learning systems. In addition, we also test the moderating effect of students' beliefs about their technology optimism on the TPB relationships, to better understand student beliefs about using weblog learning systems. Therefore, the purposes of this study are as follows: to investigate the factors that influence behavioural intention to use weblog learning systems in the education context; to develop an extended TPB model that includes how individual characteristics, technology characteristics and system quality related to weblog learning systems; to examine whether PBC, attitude towards, social influences and technology optimism moderate/predict behavioural intention to use weblog

learning systems; and to empirically assess the resulting model. To achieve these purposes, this study addresses the following research questions:

RQ1. What factors determine students' behavioural intention to use weblog learning systems for educational purposes?

RQ2. Do individual characteristics, technology characteristics and system quality affect the TPB model in weblog learning system contexts?

RQ3. How does the technology optimism mix moderate the effects of PBC, attitude towards and social influences on behavioural intention to use weblog learning systems?

This study proposes an integrated model by exploring the adoption of weblog learning as part of the theoretical framework, and makes three key contributions to the literature. First, we develop and validate a model combining the TPB and TTF in terms of students' weblog learning intention. Second, we present the moderation results associated with students' beliefs regarding their technology optimism through the TPB relationships; to our knowledge, this is the first empirical study to test the moderating role of technology optimism on students' beliefs about weblog learning. Third, we develop a new instrument to measure learner-technology fit at the individual level. In addition, from a research perspective, the research model is based on the learner perspective, and can help educational institutions to promote weblog learning among students. From a managerial aspect, the recognition of proper learner-technology fit profiles can lead to remunerative results. The next section will delineate the theoretical foundations of our conceptual model.

2. Literature review and conceptual model

This study proposes and develops a conceptual model of weblog learning intention based on the TPB and TTF, and draws from previous literature that used the two theories in a technology educational context. The model combines the TPB, TTF and technology optimism as additional predictor variables, and includes a number of individual differences as moderators.

2.1 Theory of planned behaviour (TPB)

The TPB was originally based on the efforts of the TRA. The TPB adapts the TRA theory and integrates belief, attitude, intention and behaviour into an individual's behaviour model. The TPB differs from the TRA by adding a third component, PBC. For the past decade, the TPB model has become influential and widely accepted for investigating behaviour in various contexts (Ajzen, 2011). It is a social psychological model used to investigate the relationships between certain variables and one's behavioural intention to participate in a certain type of behaviour (Ajzen, 1991). The TPB suggests that one's behaviour is predicted by intention, and the intention is mainly determined by attitude, subjective norms and PBC concerning the behaviour (Ajzen, 1991). Ajzen (1988) developed the TPB and attempted to explain individuals' behaviour under volitional and non-volitional control. The TPB is useful for testing psychological factors because it not only includes most of these psychological factors, but also helps to discover the determinants of behaviour (Armitage and Conner, 2001; Wang and Ritchie, 2010). In addition, the TPB has been widely applied across a range of disciplinary literature over the past decade, including IS usage behaviour and online learning systems (Ajzen, 1991; Armitage and Conner, 2001; Wang and Ritchie, 2010; Wu *et al.*, 2011; Yu and Yu, 2010).

2.2 Task-technology fit (TTF)

Though some prospective users use an IS and believe that the system improves their work performance, their behaviour is not necessarily voluntarily (Goodhue and Thompson, 1995). In addition, while some users assume that a system is helpful and convenient for some tasks,

the TPB model do not particularly produce strong evidence (Dishaw and Strong, 1999; Goodhue and Thompson, 1995; Nance and Straub, 1996). The TTF model is more suited to measuring the ability of an IS to assist with tasks. The TTF is a powerful model used to analyse user's adoption and use behaviour of an innovative IT designed for a specific context, and is commonly adopted to construe and predict how the fit between task requirements and technology functions positively impact the outcome of task performance and technology utilisation (Goodhue and Thompson, 1995). The TTF (Goodhue, 1995) considers the interactions between tasks, defined as actions executed by individuals to turn inputs into outputs; technologies, defined as the tools used by individuals in finishing their tasks; and capabilities. The TTF model includes: task characteristics, which impact performance; and technology characteristics (or functions), which affect the outcome (utilisation) (Goodhue and Thompson, 1995). Technology utilisation is influenced by the fit between task characteristics and technology characteristics.

2.3 Technology readiness

In the academic literature, TR has attracted considerable attention. TR was developed by Parasuraman (2000): it is a scale that combines beliefs and feelings related to technology, and determines an individual's overall predisposition to adopt new technology products and services (Ferreira *et al.*, 2014). Parasuraman used TR to measure the level of one's personality as being central to their technology acceptance (Vize *et al.*, 2013). The TR model has been applied to a variety of contexts including: online services, SNS, educational choice and healthcare services (Borrero *et al.*, 2014; Rosen *et al.*, 2003; Taylor *et al.*, 2002).

TR categorises consumers' technology adoption tendencies into one of four dimensions: optimism, innovativeness, discomfort and insecurity (Borrero *et al.*, 2014; Chung *et al.*, 2015; Ferreira *et al.*, 2014; Musah *et al.*, 2015; Parasuraman, 2000; Parasuraman and Colby, 2001; Park and Salvendy, 2012). Two of these dimensions, optimism and innovativeness, are the key drivers of TR, and lead to growth in terms of the tendency to adopt new technology. Optimism refers to a positive view about technology and a belief that it can increase control, flexibility and efficiency in life (Parasuraman, 2000; Vize *et al.*, 2013). Son and Han (2011) argue that the controllability of new technologies is essential to optimistic customers, since convenience is the most addressed advantage of accessing new technologies. Further, compared to pessimists, optimistic customers tend to focus on positive events and embrace new technology more openly. Technological advancements have resulted in students with relatively developed IT skills, and positive perceptions about IT. Moreover, students believe that weblog learning systems offer increased management control, flexibility and efficiency in learning. Based on the above discussion, it is presumed that highly optimistic students will use new learning methods more frequently and be willing to try a variety of innovative functions. Thus, this study posits that technology optimism will have a relatively significant influence on students' usage of weblog learning.

2.4 Hypotheses development

TTF provides a further explanation of the interrelationship between individual and task characteristic fitness, which impacts decisions about whether or not to use an IS. Both individual characteristics and task characteristics are essential parts of the task performance chain. For the adoption of a technology to be valid, the TTF theory argues that this technology must be voluntarily accepted and adopted by users; this requires a good fit between the technology and users, as well as a between the technology and required tasks.

Goodhue and Thompson (1995) proposed that after using an IS, users would have positive or negative feelings, and might change their expectations of and future use of the system. Lee and Lehto (2013) stated that within the TTF model, the higher the support a given technology provides for a task, the higher the perception of TTF, and the higher the

technology utilisation. Jarupathirun and Zahedi (2007) found higher TTF helps one to achieve the requirements needed for a task. Yu and Yu (2010) integrated the TTF model and the TPB model to understand learners' behaviour, perceptions and influence in terms of learner performance, which is crucial to predicting the use of electronic learning systems. They found that the TTF and TPB model constructs facilitate weblog learning performance, and offer important implications for understanding learner performance in online learning environments. To summarise, many previous researchers (Dishaw and Strong, 1999; Lee and Lehto, 2013; Yen *et al.*, 2010; Yu and Yu, 2010) have suggested that the TTF model can be expanded to other similar contexts to provide a more complete analysis of the relationship among task, technology, TTF and the use of the technology. Thus, the current study incorporates two constructs, technology characteristics and individual characteristics, which are expected to impact students' attitudes towards the system, leading to the following hypotheses:

H1. A student's individual characteristics associated with weblog learning will be positively related to the student's attitude towards the system.

H2. Perceived technology characteristics associated with weblog learning will be positively related to the student's attitude towards the system.

H3. Perceived system quality associated with weblog learning will be positively related to the student's attitude towards the system.

While the acceptance of weblog learning generally involves both technological and behavioural aspects for individual use, the TPB is not complete in its coverage of both aspects. Many studies have therefore proposed the concept of integrating TPB in a complementary manner. According to the TPB, subjective norm, PBC and attitude may influence intention, and in turn influence behaviour (Ajzen, 1991). Several hypotheses may results from using the TPB model as the base for a predictive model of weblog learning. As for the three antecedents of the intention, connections have been built between attitude and intention (Lam *et al.*, 2007; Taylor and Todd, 1995; Yu and Yu, 2010; Cheng *et al.*, 2012), social influences and intention (Borrero *et al.*, 2014; Hsu and Lin, 2008; Taylor and Todd, 1995; Venkatesh *et al.*, 2003; Teo, 2010) and PBC and intention (Armitage and Conner, 2001; Taylor and Todd, 1995; Venkatesh and Davis, 1996; Yu and Yu, 2010). Based on the above, there are three direct antecedents for determining behavioural intention: PBC, attitude towards and social influences. This study argues that there are three potential linkages between these three antecedents and behavioural intention in the weblog learning usage context, leading to the following three hypotheses:

H4. Perceived behaviour control associated with weblog learning will be positively related to the student's behavioural intention to use the system.

H5. A student's attitude towards weblog learning will be positively related to the student's behavioural intention to use the system.

H6. A student's social influences associated with weblog learning will be positively related to the student's behavioural intention to use the system.

Technology optimism proposes that technology brings increased control, flexibility and efficiency (Borrero *et al.*, 2014). Students nowadays are considered digital natives, and most of them they have positive opinions towards technology usage (Lewis and Mayes, 2014; Musah *et al.*, 2015). For instance, the application of weblog learning is common, and it also improves learning performance. This shows that students have relatively high level of technology optimism. However, when examining students' intention to use weblog learning, the reasons for usage should not be the sole focus; instead, the influence of their level of

technology optimism should be emphasised even more. The level of technology optimism can influence students' opinions about weblog learning, such as having a positive attitude towards weblog learning or thinking weblog learning will be a good learning approach. It is assumed that technology optimism is a key moderator in the relationship among PBC, attitude towards, social influences and behavioural intention. This study expects respondents with high or low technology optimism to have different belief–intention relationships. On this basis, the following hypotheses are proposed:

H7. The relationship between perceived behavioural control and behavioural intention is moderated by the level of technology optimism: that is, the relationship is weaker under conditions of low technology optimism and stronger under conditions of high technology optimism.

H8. The relationship between attitude towards and behavioural intention is moderated by the level of technology optimism: that is, the relationship is weaker under conditions of low technology optimism and stronger under conditions of high technology optimism.

H9. The relationship between social influences and behavioural intention is moderated by the level of technology optimism: that is, the relationship is weaker under conditions of low technology optimism and stronger under conditions of high technology optimism.

Based on the above, a conceptualised structural model demonstrating the moderating role of technology optimism is presented in Figure 1.

3. Methods

3.1 Measure development and validation

This study used questionnaires and interviews as the main instruments for data collection. The two-part questionnaire included 41 questions: the first part focussed on respondents' basic information using nominal and ratio scales; the second part examined respondents' perceptions of TTF, TPB, social influences and technology optimism. Table I denotes the constructs used in this study as well as their origins. This instrument included 38 items to

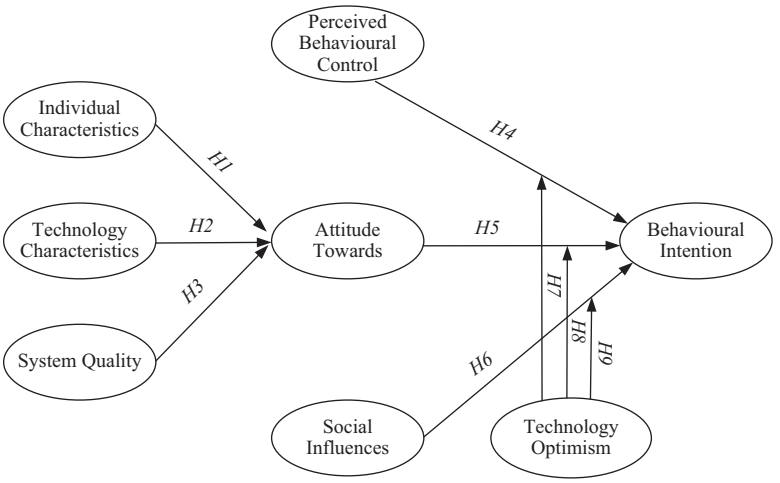


Figure 1.
The conceptual of structural equation model

Construct	Definition
Individual characteristics	Users' need for work or coursework using the weblog learning system
Technology characteristics	Users' understanding of the features of the weblog learning system
System quality	The support quality that users receive from the weblog learning system
Technology optimism	A positive attitude towards technology and a belief in increased control, flexibility and efficiency in one's life
Perceived behavioural control	An individual's perceived ease or difficulty of using the weblog learning system
Social influences	An individual's perception about using the weblog learning system that is influenced by the judgement of significant others (e.g. parents, friends, teachers, etc.)
Attitude towards	An individual's positive or negative feelings or appraisals about using the weblog learning system
Behavioural intention	An indication of an individual's readiness to use the weblog learning system

Table I.
Construct definitions

measure the seven constructs of the structural model. Perception measurement items for each construct asked respondents to respond to each statement on a Likert-type scale anchored by 1 and 7, where 1 = strongly disagree and 7 = strongly agree.

In this study, group interviews were added after the analysis in order to gather more in-depth data on student perceptions. The purpose of the interviews was to further understand students' perceptions and experiences associated with academic weblogs. The interviewees were either volunteers or chosen purposefully. First, a group e-mail was sent to the participants asking for volunteers. Based on the replies, individual e-mails were sent out to selected students in order to ensure variance with respect to gender, level of participation with academic weblogs and past experience with weblogs. Finally, nine business and management students (four males and five females) were recruited. The interview questions were mostly semi-structured, including both general queries and customised questions based on students' specific use of weblogs. Each interview lasted 30–40 min and was recorded and transcribed for later analysis.

According to the TPB perspective and following Ajzen (1988), all items were adapted to specifically measure respondents' weblog learning systems adoption behaviours. The instruments were developed after a thorough review of previous research specifically pertaining to the TTF, TPB and technology optimism in theory and in practice. Measurement items were modified to confirm to the weblog learning system adoption context. Our scale development followed the recommendations of Straub (1989) and standard psychometric scale development procedures (DeVellis, 2003; MacKenzie *et al.*, 2011). The TTF constructs (technological characteristics, individual characteristics and system quality) were adapted from several studies (Goodhue, 1995; Goodhue and Thompson, 1995; Lee and Lehto, 2013; Närman *et al.*, 2012; Yen *et al.*, 2010; Yu and Yu, 2010). The final instrument was composed of 14 statements pertaining to technological characteristics (four items), individual characteristics (five items) and systems quality (five items). The TPB constructs (attitude, social influences, PBC, behavioural intention) were from several other researchers (Ajzen, 1991, 2011; Borrero *et al.*, 2014; Hsu and Lin, 2008; Venkatesh *et al.*, 2003; Wu *et al.*, 2011; Yu and Yu, 2010). The attitude measure had five items, social influences had three items, perceived behavioural control had four items and behavioural intention had five items. Finally, learners' technology optimism behaviour measurements were adapted from several previous researchers (Borrero *et al.*, 2014; Chung *et al.*, 2015; Ferreira *et al.*, 2014; Parasuraman, 2000; Vize *et al.*, 2013, Lewis and Mayes, 2014) and measured using seven items.

To ensure that the survey questionnaires were worded in a concise and understandable manner, an initial pilot study was carried out. The initial questionnaires were administered to 76 students who reported that they had used and were familiar with weblog learning systems. The revised questionnaire was evaluated for readability, ease of understanding and formatting issues prior to the actual test. A reliability analysis (Cronbach's α) was also performed to test the reliability and internal consistency of each of the 38 attributes measured. Based on participant feedback, seven items were eliminated due to substantive semantic overlap with other items. The Cronbach's α reliability scores for the remaining 31 attributes ranged from 0.810 for individual characteristics to 0.931 for attitude towards and were all above the minimum value of 0.6, that is, considered acceptable as an indication of reliability (Hair *et al.*, 2010). This implies that the scales used in this study were satisfactory in terms of measuring the constructs of interest. Based on the results of the pilot sample, minor modifications were made to the survey design.

3.2 Sample and descriptive statistics

This study use a Moodle learning system integrated with blogs that allowed users to reflect on and enhance their learning practices. This system allowed learners and instructors to connect both socially and academically. Students self-managed their learning process using the Moodle system across four phases: getting ready (preparing oneself to use the tool in the activity), setting out (beginning to use the tool in the activity), carrying on (the processes of doing a particular action within the activity) and finishing off (process tool use captures the highest point of the action). The current study aimed to determine specific factors that influence weblog learning systems. To fulfil these objectives, as noted previously, this study employed online and interview surveys. For the surveys, respondents identified their absolute usage of the weblog learning system; next, potential participants received invitation e-mails to answer a series of questions through a URL linked to a web-based survey form. Participation in the study was completely voluntary, but was limited to subjects over the age of 18 with previous weblog learning experience. Subjects consisted of undergraduate and graduate students at three private universities in the south of Taiwan. In total, 380 usable responses were obtained. The average age of the respondents was 20.43 years, and 58.9 per cent were female. The respondents had used weblog learning systems for an average of 3.23 years, and had used them to complete an average of 3.34 courses.

4. Results

Depicting a model containing moderators with PLS differs from a traditional representation of that research model. With a PLS model, the moderator, in this case the construct technology optimism, is shown as an independent variable with a direct path to behavioural intention. These interaction measurement variables are based on Chin *et al.*'s (2003) suggestion to multiply every indicator in the moderator by every indicator in the independent variable. Conceptually, the interaction constructs (technology optimism × perceived behavioural control, technology optimism × attitude towards and technology optimism × social influences) are depicted as having a direct path to behavioural intention.

4.1 Measurement model evaluation

Internal consistency was assessed by the Cronbach's α and composite reliability coefficients. The Cronbach's α coefficients ranged from 0.830 to 0.995, and the results presented in Table II attest to the high internal consistency of the instrument, as all values exceeded the suggested 0.70 level for scale robustness (Hair *et al.*, 2010). Composite reliability means that a set of latent construct indicators are consistent in their

Construct	AVE	Composite reliability	Cronbach's α	Discriminant validity	ICs	TCs	SQ	TO	PBC	SI	ATT	BI
Individual characteristics (ICs)	0.665	0.888	0.830	2.175	*0.815*							
Technology characteristics (TCs)	0.711	0.908	0.865	2.150	0.551	*0.843*						
System quality (SQ)	0.981	0.996	0.995	25.815	0.185	0.157	*0.990*					
Technology optimism (TO)	0.749	0.900	0.833	1.950	0.542	0.575	0.125	*0.865*				
Perceived behavioural control (PBC)	0.975	0.992	0.987	46.545	0.108	0.110	0.020	0.143	*0.987*			
Social influences (SI)	0.803	0.924	0.876	2.298	0.464	0.455	0.150	0.550	0.093	*0.896*		
Attitude towards (ATT)	0.688	0.898	0.849	1.407	0.553	0.514	0.169	0.591	0.104	0.576	*0.829*	
Behavioural intention (BI)	0.633	0.896	0.855	1.295	0.502	0.480	0.195	0.620	0.145	0.591	0.699	*0.796*

Notes: Discriminant validity = AVE/(Correlation)^2, where (Correlation)^2 = highest (Correlation)^2 between factors of interest and remaining factors

Table II.
Constructs reliability result

measurement (Hair *et al.*, 2010; Jöreskog and Sörbom, 2005). The composite reliability coefficients in this study ranged from 0.888 to 0.996, which all exceed the benchmark of 0.60 (Fornell and Larcker, 1981). Convergent and discriminant validity were evaluated by calculating the average variance extracted (AVE) for each construct. Validity is established if the shared variance accounts for 0.50 or more of the total variance. In the current study, the AVE values ranged for 0.633 to 0.981. Discriminant validity is evident when the AVE for each construct is greater than the squared correlation coefficients between that construct and any other construct in the model (Fornell and Larcker, 1981). As shown in Table II, overall, the constructs demonstrated satisfactory convergent validity and discriminant validity.

4.2 Interpretation of structural model testing

The majority of research on user involvement and weblog learning success has looked at the impact of user involvement on user attitudes and ultimately on user commitment to utilise weblogs. This research combined the TTF, TPB and individual technology optimism, and proposed a better hybrid technology behavioural intention model to explain students' usage behaviour of online learning course websites. In sum, two important technology usage models, the TTF and TPB, were included to understand their predictive power on learner use of weblog learning systems among Taiwan students. The standardized solution estimated by the Smart PLS M3 programme was used to interpret the structural relation results. Figure 2 depicts that the R^2 and path coefficients (loading and significance) can support the hypothesised model. The paths specified in the integrated model were all significant with the direct and indirect effects of attitude, accounting for 61.5 per cent of the variance in behavioural intention. The proposed model explained a significant amount of variation regarding the endogenous variables (i.e. more than 35 per cent on average). Endogenous variables showed good explanatory power of variation, which indicates reasonable model stability and robustness. The statistical significance of each path was judged according to a critical value, which is also referred to as the 0.05 level of significance. All six causal paths, one moderator effect path and three interaction effect paths were specified in the proposed model, and eight were found to be statistically significant.

The components of individual characteristics (*H1*) and technology characteristics (*H2*) were both important antecedents of attitude ($\beta = 0.377$ and 0.297), whereas system quality (*H3*) was non-significant in terms of attitude ($\beta = 0.052$, $p > 0.05$). Therefore, *H1* and *H2* are

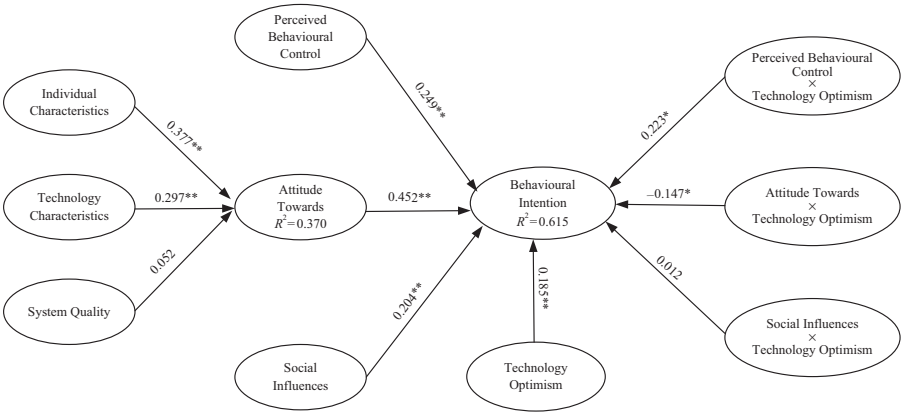


Figure 2.
Path coefficients for the online learning system behavioural intention model

Notes: *$p<0.05$; **$p<0.01$

supported, but *H3* is not. The construct of perceived behavioural control (*H4*), attitude towards (*H5*) and social influences (*H6*) were significant determinants of behavioural intention ($\beta = 0.249$, 0.452 and 0.204, respectively), in support of *H4–H6*. The direct effect of technology optimism on behavioural intention ($\beta = 0.185$) and the interaction effect of technology optimism and perceived behavioural control (*H7*) and attitude towards (*H8*) behavioural intention ($\beta = 0.223$ and -0.147) were significant, whereas social influences (*H9*) was non-significant in terms of behavioural intention ($\beta = 0.012$, $p > 0.05$). Therefore, *H7* and *H8* are supported, but *H9* is not.

We also hypothesised that positive technology optimism would moderate the impact of social influences; however, no support was found for this, and the predictor was also insignificant in the PLS analysis. That is, the main effect of social influences and the moderating impact of positive technology optimism were not supported. Moreover, system quality was not found to be an antecedent of attitude towards weblog learning. The research model weaknesses regarding users' attitudes may be primarily attributable to the lack of explicit inclusion of individual characteristics, as well as the degree to which the weblog learning system met users' perceived system quality needs. This weakness may be especially apparent in our data because the tool function and system icons were difficult to personalise.

As shown in Figure 3, the interaction between attitude towards and technology optimism was significant. Under the interaction of technology optimism, as students' weblog learning attitude increased, their behavioural intention of usage decreased. Nowadays, students know how to make good use of weblogs to enhance their learning effectiveness. However, after reaching a certain level of understanding, they consider weblog effectiveness to be limited, and no longer spend as much time using them to learn, resulting in weakened weblog learning behaviours. As for PBC, under the interaction of technology optimism (as see Figure 4), as PBC towards weblog learning increases, so does behavioural intention to use weblog learning; further, technology optimism strengthens the effect of PBC. When students recognise that weblog learning can be a very useful way to learn and enhance their academic efficiency, their technology optimism tends to increase. Consequently, they actively seek out solutions or resources to overcome bottlenecks associated with weblog learning, and continue to use it. Moreover, even though social influence affected the behavioural intention of weblog learning, the result was not significant, as shown in Figure 5. Through interviewing the students, we found that the main reason was that students all listen to and adopt recommendations made by classmates or friends when dealing with a new type of IT, and will use weblogs to search, read and learn. However, students worried about the reliability of weblog information providers' information, resulting in an interaction effect between social influence and technology optimism that did not significantly influence the usage intention of weblog learning.

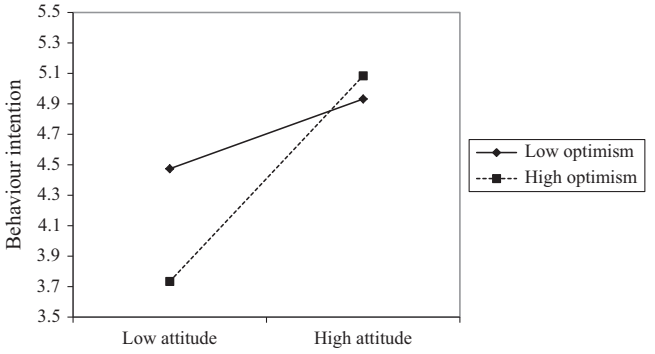


Figure 3.
Technology optimism and attitude towards: the moderating role of behavioural

4.3 Common method bias

Common method bias is a potential threat to internal validity, particularly when using similar scales that collect responses in a single survey setting. The threat is high if a single factor can account for the majority of covariance between the independent and dependent variables (Podsakoff *et al.*, 2003). We ran a factor analysis (e.g. Harman's one-factor test) to demonstrate that no single factor loaded on all measures, and the results support this. Therefore, common method bias was not considered to be a problem in this study.

5. Discussion

The application of weblogs in teaching has promising potential to improve on the weaknesses of traditional teaching approaches while becoming a new channel of learning. This study using a Moodle learning system integrated with a blog allowed learners and instructors to connect both socially and academically. The learning system had several primary functions, each aligned with the constituent processes of weblog learning (assessing, planning, implementing, monitoring and evaluating). The TPB is the most widely used and most influential theory for investigating human behaviours (Ajzen, 2011). However, this theory does not include variables such as individual technological skills and personal needs related to the pre-variables that influence behaviours. Some variables have been suggested to extend the original TPB due to the importance of individual and technology characteristics (Kwon and Wen, 2010). External variables have been included based on the specific characteristics of the technology, such as individual, task and system characteristics. The TTF model is an attempt to resolve the limitations of the TPB.

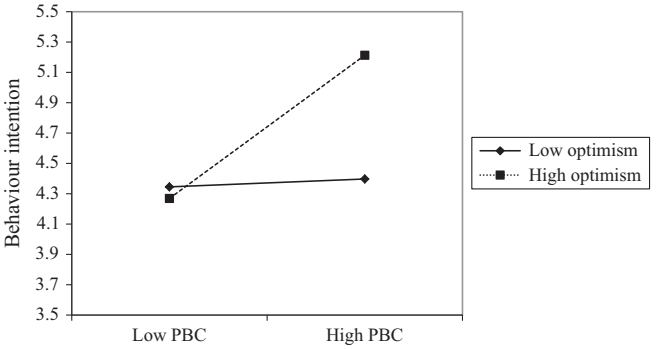


Figure 4.
Technology optimism and PBC: the moderating role of behavioural

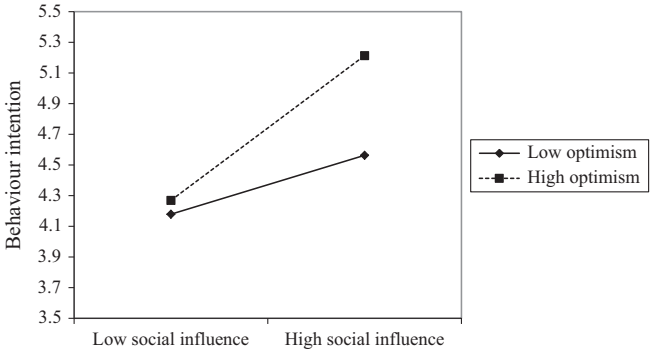


Figure 5.
Technology optimism and social influence: the moderating role of behavioural

In addition, various studies have attempted to understand users' technology optimism and predict their behavioural intention. The "technology optimism" as an important dimension of TR, contributing to an individual's increased propensity to adopt new system or technology. The current study integrates viewpoints of TPB and TTF and proposes a causal model between blog learning behaviour and behavioural intention. For students who are optimistic about science and technology and possess high self-efficacy when using IT, the level of technology optimism affects their willingness to use IS. Accordingly, this study includes the level of technology optimism as a moderator to explore whether it impacts perceived behavioural control, attitude and social influences on behavioural intention related to weblog learning. Through collection and analysis of empirical data, we construct and test a structural equation model (SEM). In addition, nine business and management students (four males and five females) participated in interviews to gain a deeper understanding of students' perceptions and experiences with using academic weblogs. The following describes the main research findings.

First, based on the explanations offered above, this study considers the pre-variables of TTF as external variables of TPB in order to explore students' attitudes about using weblogs so as to measure changes between weblog technology and students' personal motivation to use them (Dishaw and Strong, 1999; Taylor and Todd, 1995; Yen *et al.*, 2010; Yu and Yu, 2010). Personal characteristics and technology characteristics affect students' attitudes towards weblog learning, whereas system quality did not have a significant impact. According to previous studies (Wang *et al.*, 2016; Yu and Yu, 2010), students nowadays utilise blogs to improve their study effectiveness. In higher education, particularly business and management education, blogs help to increase intellectual interactions among students. The current study focusses on educational weblog learning systems, where they can post content about assignments and discuss their thoughts on course materials. The results of the nine interviews show that when students adopt weblog learning, they care more about the thoroughness, meaning and value of the content. In addition, the interviewees indicated that they believe weblog learning systems to be solid and stable. The system used in this study allowed for increased interaction for all parties, and rapid responses to posted opinions. The hardware devices used by learners were sufficient to assist in their learning. However, system quality did not affect students' attitudes towards weblog learning. In summary, the research model can help researchers and students to clarify the external variables that influence individual attitudes towards weblog learning.

Attitude is the most important factor that affects students' behavioural intentions towards weblog learning, and in this study it had a significantly positive effect, which is consistent with previous results (Lam *et al.*, 2007; Taylor and Todd, 1995; Yu and Yu, 2010). The Moodle learning system used in this study required that students self-manage their learning process, and use the system for the four phases: getting ready, setting out, carrying on and finishing off. When students believed that weblogs would help them better control the learning activities and manage their time, their behavioural intention was higher, which in turn enhanced their understanding of weblog learning, and increased their usage intention. Further, perceived behavioural control had a significantly positive effect on behavioural intention, similar to results obtained in previous studies (Armitage and Conner, 2001; Taylor and Todd, 1995; Venkatesh and Davis, 1996; Yu and Yu, 2010). The weblog learning environment is better for students due to the opportunities for interactive learning using the internet. Students used weblogs to discuss classwork details and share important information, leading to knowledge growth. When students encountered a problem with using weblogs, they used the internet to find a solution, and thereby reduced learning obstacles. Moreover, when students' classmates or friends also used weblogs, behavioural intention of usage was strengthened through social networking, suggesting that social

influences can enhance weblog behavioural intention of usage, as noted by previous researchers (Borrero *et al.*, 2014; Hsu and Lin, 2008; Teo, 2010).

In recent years, many studies have attempted to understand users' TR to effectively predict their behaviours. Technology optimism is an important dimension of TR, contributing to an individual's increased propensity to adopt new systems or technology. This study was intended to better understand students' usage of weblog learning and their perceptions on how weblogs can improve study efficiency and convenience through an investigation of the moderating effect of students' beliefs about their technology optimism on TPB relationships. To our knowledge, this is the first empirical study to test the moderating role of technology optimism in this context. The research results pertaining to the interactions among attitude, perceived behavioural control, social influence and technology optimism indicate that attitude, perceived behavioural control and technology optimism have a significant influence on behavioural intention.

Finally, for students with better weblog learning attitudes, the higher their technology optimism, the weaker the relationship between attitude and behavioural intention; for students with better weblog learning perceived behavioural control, the higher their technology optimism, the stronger the relationship between attitude and behavioural intention.

6. Conclusions

This study combines the literature on the TPB, TTF, technology optimism and weblog learning system use behavioural intention and proposes a moderating model of the weblog learning system use behavioural intention of business and management students at southern Taiwan universities through a SEM to test our empirical model. This is the first empirical study to test the moderating role of technology optimism on understanding students' beliefs regarding the use of weblog learning systems. The empirical results lead to three important conclusions: behavioural intention of the use of weblog learning systems is dependent on the attitude towards it, social influences, and perceived behaviour control; students' attitudes towards the system are dependent on individual and technological characteristics; and students' technology optimism significantly moderates their attitudes towards these systems, as well as perceived behaviour control factors. Further, our research model is based on the learner perspective and can help educational institutions to promote weblog learning among students. The conceptual model presented in this study can serve as an activator for research on weblog learning and encourage further explorations in this area.

7. Research limitations and future research directions

This paper focusses on modelling students' usage intention of weblog learning through an empirical study, in order to better understand how technology optimism influences perceived behavioural control, attitude towards and the moderating effect of social influence on behavioural intention to use weblog learning. However, we did not further analyse or explore the relationship between usage intention and actual usage behaviour. In addition, college students served as the research sample for the questionnaire survey, and individual technology optimism as the moderator. We did not specifically separate the participants into high and low technology optimism groups to test model differences, which future researchers can consider. This study also explored weblog learning associated with SNSs. In order to increase external validity, future researchers can examine SNSs so as to better understand relevant behavioural usage intentions. Lastly, one moderator in this study was individual technology characteristics. Other variables may also moderate behavioural usage intention, such as knowledge content quality and knowledge trust. Future studies can consider including these variables as moderators.

References

Ajzen, I. (1988), *Attitudes, Personality, and Behavior*, Open University Press and Dorsey Press, Milton-Keynes and Chicago, IL.

Ajzen, I. (1991), "The theory of planned behavior", *Organizational Behavior and Human Decision Processes*, Vol. 50 No. 2, pp. 179-211.

Ajzen, I. (2011), "The theory of planned behavior", *Handbook of Theories of Social Psychology*, Vol. 1, Lawrence Erlbaum, New York, NY, pp. 438-459.

Armitage, C.J. and Conner, M. (2001), "Efficacy of the theory of planned behaviour: a meta-analytic review", *British Journal of Social Psychology*, Vol. 40 No. 4, pp. 471-499.

Borrero, J.D., Yousafzai, S.Y., Javed, U. and Page, K.L. (2014), "Expressive participation in Internet social movements: testing the moderating effect of technology readiness and sex on student SNS use", *Computers in Human Behavior*, Vol. 30, pp. 39-49.

Chen, C.P., La, H.M. and Ho, C.Y. (2015), "Why do teachers continue to use teaching blogs? The roles of perceived voluntariness and habit", *Computers & Education*, Vol. 82, pp. 236-249.

Cheng, P.Y., Hsu, P.K. and Chiou, W.B. (2012), "Undergraduates' intentions to take examinations for professional certification: examinations of four competing models", *Asia Pacific Education Review*, Vol. 13 No. 4, pp. 691-700.

Chin, W.W., Marcolin, B. and Newsted, P. (2003), "A partial least squares latent variable modeling approach for measuring interaction effects: results from a Monte Carlo simulation study and an electronic-mail emotion/adoption study", *Information Systems Research*, Vol. 14 No. 2, pp. 189-217.

Chung, N., Han, H. and Joun, Y. (2015), "Tourists' intention to visit destination: role of augmented reality applications for heritage site", *Computers in Human Behavior*, Vol. 50, pp. 588-599.

Chwo, G.S.M. (2015), "Empowering EIL learning with a Web 2.0 resource: an initial finding from the cross campus Storybird feedback study", *Computers & Education*, Vol. 84, pp. 1-7.

Deng, L. and Yuen, A.H.K. (2011), "Towards a framework for educational affordances of blogs", *Computers & Education*, Vol. 56 No. 2, pp. 441-451.

DeVellis, R.F. (2003), *Scale Development: Theory and Applications*, 2nd ed., Sage Publications, Thousand Oaks, CA.

Dishaw, T. and Strong, M. (1999), "Extending the technology acceptance model with task-technology fit constructs", *Information & Management*, Vol. 36 No. 1, pp. 9-21.

Ferreira, J.B., de Rocha, A. and da Silva, J.F. (2014), "Impacts of technology readiness on emotions and cognition in Brazil", *Journal of Business Research*, Vol. 67 No. 5, pp. 865-873.

Fornell, C. and Larcker, D.F. (1981), "Evaluating structural equation models with unobservable variables and measurement error", *Journal of Marketing Research*, Vol. 18 No. 1, pp. 39-50.

Freeman, W. and Brett, C. (2012), "Prompting authentic blogging practice in an online graduate course", *Computers & Education*, Vol. 59 No. 3, pp. 1032-1041.

Furukawa, T., Matsuo, Y., Matsuzawa, T., Takeda, M. and Uchiyama, K. (2006), "Users' behavioral analysis on weblogs", available at: http://aaaipress.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-007.pdf (accessed 10 February 2011).

Goodhue, D.L. and Thompson, R.L. (1995), "Task-technology fit and individual performance", *MIS Quarterly*, Vol. 19 No. 2, pp. 213-236.

Goodhue, L. (1995), "Understanding user evaluations of information systems", *Management Science*, Vol. 41 No. 12, pp. 1827-1844.

Hair, F. Jr, Black, W.C., Babin, B.J. and Anderson, R.E. (2010), *Multivariate Data Analysis: A Global Perspective*, 7th ed., MacMillan, New York, NY.

Hsu, C.L. and Lin, J.C.C. (2008), "Acceptance of blog usage: the roles of technology acceptance, social influence and knowledge sharing motivation", *Information & Management*, Vol. 45 No. 1, pp. 65-74.

Jarupathirun, S. and Zahedi, F.M. (2007), "Exploring the influence of perceptual factors in the success of web-based spatial DSS", *Decision Support Systems*, Vol. 43 No. 3, pp. 933-951.

Jimoyiannis, A. and Angelaina, S. (2012), "Towards an analysis framework for investigating students' engagement and learning in educational blogs", *Journal of Computer Assisted Learning*, Vol. 28 No. 3, pp. 222-234.

Jöreskog, K.G. and Sörbom, D. (2005), *LISREL 8.72: A Guide to the Program and Applications*, 3rd ed., Scientific Software International, Chicago, IL.

Kang, I., Bonk, C.J. and Kim, M.C. (2011), "A case study of blog-based learning in Korea: technology becomes pedagogy", *Internet and Higher Education*, Vol. 14 No. 4, pp. 227-235.

Kerawalla, L., Minocha, S., Kirkup, G. and Conole, G. (2009), "An empirically grounded framework to guide blogging in higher education", *Journal of Computer Assisted Learning*, Vol. 25 No. 1, pp. 31-42.

Kim, H.N. (2008), "The phenomenon of blogs and theoretical model of blog use in educational contexts", *Computers & Education*, Vol. 51 No. 3, pp. 1342-1352.

Kwon, O. and Wen, Y. (2010), "An empirical study of the factors affecting social network service use", *Computers in Human Behavior*, Vol. 26 No. 2, pp. 254-263.

Ladyshewsky, R.K. and Gardner, P. (2008), "Peer assisted learning and blogging: a strategy to promote reflective practice during clinical fieldwork", *Australasian Journal of Educational Technology*, Vol. 24 No. 3, pp. 241-257.

Lam, T., Cho, V. and Qu, H. (2007), "A study of hotel employee behavioral intentions towards adoption of information technology", *International Journal of Hospitality Management*, Vol. 26 No. 1, pp. 49-65.

Larsen, T.J., Sørebø, A.M. and Sørebø, Ø. (2009), "The role of task-technology fit as users' motivation to continue information system use", *Computers in Human Behavior*, Vol. 25 No. 3, pp. 778-784.

Lee, D.Y. and Lehto, M.R. (2013), "User acceptance of YouTube for procedural learning: an extension of the technology acceptance model", *Computers and Education*, Vol. 61, pp. 193-208.

Lee, J. and Bonk, C.J. (2016), "Social network analysis of peer relationships and online interactions in a blended class using blogs", *Internet and Higher Education*, Vol. 28, pp. 35-44.

Lewis, J.R. and Mayes, D.K. (2014), "Development and psychometric evaluation of the emotional metric outcomes (EMO) questionnaire", *International Journal of Human-Computer Interaction*, Vol. 30 No. 9, pp. 685-702.

Lin, W.-S. (2012), "Perceived fit and satisfaction on web learning performance: IS continuance intention and task-technology fit perspectives", *International Journal of Human-Computer Studies*, Vol. 70 No. 7, pp. 498-507.

Lu, H.P. and Hsiao, K.L. (2009), "Gender differences in reasons for frequent blog posting", *Online Information Review*, Vol. 33 No. 1, pp. 135-156.

MacKenzie, S.B., Podsakoff, P.M. and Podsakoff, N.P. (2011), "Construct measurement and validation procedures in MIS and behavioral research: integrating new and existing techniques", *MIS Quarterly*, Vol. 35 No. 2, pp. 293-334.

Ma, C.M., Chao, C.M. and Cheng, B.W. (2013), "Integrating technology acceptance model and task -technology fit into blended e-learning system", *Journal of Applied Sciences*, Vol. 13 No. 5, pp. 736-742.

Mohammadyari, S. and Singh, H. (2015), "Understanding the effect of e-learning on individual performance: the role of digital literacy", *Computers & Education*, Vol. 82, pp. 11-25.

Musah, M.B., Ali, H.B.M., Al-Hudawi, S.H.V., Tahir, L.M., Daud, K.B. and Hamdan, A.R. (2015), "Determinants of students' outcome: a full-fledged structural equation modelling approach", *Asia Pacific Education Review*, Vol. 16 No. 4, pp. 579-589.

Nance, W.D. and Straub, D.W. (1996), "An investigation of task/technology fit and information technology choices in knowledge work", *Journal of Information Technology Management*, Vol. 7 Nos 3/4, pp. 1-14.

Närman, P., Holm, H., Höök, D., Honeth, N. and Johnson, P. (2012), "Using enterprise architecture and technology adoption models to predict application usage", *The Journal of Systems and Software*, Vol. 85 No. 8, pp. 1953-1967.

Parasuraman, A. (2000), "Technology readiness index (TRI): a multiple-item scale to measure readiness to embrace new technologies", *Journal of Service Research*, Vol. 2 No. 4, pp. 307-320.

Parasuraman, A. and Colby, C. (2001), *Techno-Ready Marketing: How and why your Customers Adopt Technology*, The Free Press, New York, NY.

Park, T. and Salvendy, G. (2012), "Emotional factors in advertising via mobile phones", *International Journal of Human-Computer Interaction*, Vol. 28 No. 9, pp. 597-612.

Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y. and Podsakoff, N.P. (2003), "Common method biases in behavioral research: a critical review of the literature and recommended remedies", *Journal of Applied Psychology*, Vol. 88 No. 5, pp. 879-903.

Robertson, J. (2011), "The educational affordances of blogs for self-directed learning", *Computers & Education*, Vol. 57 No. 2, pp. 1628-1644.

Rosen, J., Mittal, V., Mulsant, B., Degenholtz, H., Castle, N. and Fox, D. (2003), "Educating families of nursing home residents: a pilot system using a computer based system", *Journal of the American Medical Directors Association*, Vol. 4 No. 3, pp. 128-134.

Son, M. and Han, K. (2011), "Beyond the technology adoption: technology readiness effects on post-adoption behavior", *Journal of Business Research*, Vol. 64 No. 11, pp. 1178-1182.

Straub, D.W. (1989), "Validating instruments in MIS research", *MIS Quarterly*, Vol. 13 No. 2, pp. 147-169.

Taylor, A., Celuch, K. and Goodwin, S. (2002), "Technology readiness in the E-insurance industry: an exploratory investigation and development of an agent technology E-consumption model", *Journal of Insurance Issues*, Vol. 25 No. 2, pp. 142-165.

Taylor, S. and Todd, P.A. (1995), "Understanding information technology usage: a test of competing models", *Information Systems Research*, Vol. 6 No. 2, pp. 144-176.

Teo, T. (2010), "Examining the influence of subjective norm and facilitating conditions on the intention to use technology among pre-service teachers: a structural equation modeling of an extended technology acceptance model", *Asia Pacific Education Review*, Vol. 11 No. 2, pp. 253-262.

Thelwall, M. and Hasler, L. (2007), "Blog search engines", *Online Information Review*, Vol. 31 No. 4, pp. 467-479.

Venkatesh, V. and Davis, F.D. (1996), "A model of the antecedents of perceived ease of use: development and test", *Decision Sciences Journal*, Vol. 27 No. 3, pp. 451-481.

Venkatesh, V., Morris, M.G., Davis, G.B. and Davis, F.D. (2003), "User acceptance of information technology: toward an unified view", *MIS Quarterly*, Vol. 27 No. 3, pp. 425-478.

Vize, R., Coughlan, J., Kennedy, A. and Ellis-Chadwick, F. (2013), "Technology readiness in a B2B online retail context: an examination of antecedents and outcomes", *Industrial Marketing Management*, Vol. 42 No. 6, pp. 909-918.

Wang, J. and Ritchie, B.W. (2010), "A theoretical model for strategic crisis planning: factors influencing crisis planning in the hotel industry", *International Journal of Tourism Policy*, Vol. 3 No. 4, pp. 297-317.

Wang, K.T., Huang, Y.M., Jeng, Y.L. and Wang, T.I. (2008), "A blog-based dynamic learning map", *Computers & Education*, Vol. 51 No. 1, pp. 262-278.

Wang, Y.-S., Li, C.-R., Yeh, C.-H., Cheng, S.-T., Chiou, C.-C., Tang, Y.-C. and Tang, T.-I. (2016), "A conceptual model for assessing blog-based learning system success in the context of business education", *The International Journal of Management Education*, Vol. 14 No. 3, pp. 379-387.

Wu, I.L., Li, J.Y. and Fu, C.Y. (2011), "The adoption of mobile healthcare by hospital's professionals: an integrative perspective", *Decision Support Systems*, Vol. 51 No. 3, pp. 587-596.

Yen, D.C., Wu, C.S., Cheng, F.F. and Huang, Y.W. (2010), "Determinants of users' intention to adopt wireless technology: an empirical study by integrating TTF with TAM", *Computers in Human Behavior*, Vol. 26 No. 5, pp. 906-915.

Yu, T.K. and Yu, T.Y. (2010), "Modelling the factors that affect individuals' utilization of online learning systems: an empirical study combining the task technology fit model with the theory of planned behavior", *British Journal of Educational Technology*, Vol. 41 No. 6, pp. 1003-1017.

Yu, T.K., Lu, L.C. and Liu, T.F. (2010), "Exploring factors that influence knowledge sharing behavior via weblogs", *Computers in Human Behavior*, Vol. 26 No. 1, pp. 32-41.

Further reading

Davis, F.D. (1989), "Perceived usefulness, perceived ease of use, and user acceptance of information technology", *MIS Quarterly*, Vol. 13 No. 3, pp. 319-340.

Corresponding author
Tai-Kuei Yu can be contacted at: yutk2000@gmail.com

Volume **43** Number **1 2019**

Online Information Review

Number 1

www.emeraldinsight.com/loi/oir

emerald PUBLISHING