Analysing the features of negative sentiment tweets

Ling Zhang
Department of Management, Wuhan University of Science and Technology, Wuhan, Hubei, China

Wei Dong
Department of Education, Tianjin University, Tianjin, China, and

Xiangming Mu
Department of Information Studies, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA

Abstract
Purpose – This paper aims to address the challenge of analysing the features of negative sentiment tweets. The method adopted in this paper elucidates the classification of social network documents and paves the way for sentiment analysis of tweets in further research.

Design/methodology/approach – This study classifies negative tweets and analyses their features.

Findings – Through negative tweet content analysis, tweets are divided into ten topics. Many related words and negative words were found. Some indicators of negative word use could reflect the degree to which users release negative emotions: part of speech, the density and frequency of negative words and negative word distribution. Furthermore, the distribution of negative words obeys Zipf’s law.

Research limitations/implications – This study manually analysed only a small sample of negative tweets.

Practical implications – The research explored how many categories of negative sentiment tweets there are on Twitter. Related words are helpful to construct an ontology of tweets, which helps people with information retrieval in a fixed research area. The analysis of extracted negative words determined the features of negative tweets, which is useful to detect the polarity of tweets by machine learning method.

Originality/value – The research provides an initial exploration of a negative document classification method and classifies the negative tweets into ten topics. By analysing the features of negative tweets, related words, negative words, the density of negative words, etc. are presented. This work is the first step to extend Plutchik’s emotion wheel theory into social media data analysis by constructing filed specific thesauri, referred to as local sentimental thesauri.

Keywords Sentiment analysis, Twitter, Features, Negative sentiment tweets, Topic classification

Paper type Research paper

Introduction
The emergence of Web 2.0 has significantly changed the way users perceive the internet. Contrary to the first generation of websites, on which users could only passively view content, Web 2.0 users are encouraged to participate and collaborate, forming virtual online communities. Microblogs, such as Twitter, are one of the popular Web 2.0 applications and

The research was sponsored in part by the National Social Science Fund Project “Study on dynamic optimisation mechanism of information diffusion in social networks”, Agreement Number 15CTQ029.
services. Such applications have evolved into practical means for sharing opinions on almost all aspects of everyday life.

The main difference between microblogs and traditional blogs is the strict constraint on content size (Kaplan and Haenlein, 2011). For example, Twitter is a popular microblogging service through which users send and receive text-based posts, known as tweets, consisting of up to 140 characters. Twitter was created in 2006 and reports approximately 200 million active users, posting 400 million tweets per day (Ritter et al., 2011). With its rapidly increasing popularity, Twitter is important in people’s daily life. Consequently, microblogging websites have become rich data sources for opinion mining and sentiment analysis (Kontopoulos et al., 2013). Twitter users can post opinions, experiences and queries on any chosen topic, and especially emotions, through commenting on sports competitions and entertainment programmes, sharing opinions on politics, shopping experiences, etc., all via electronic means.

In this paper, tweet data were collected over one month and were classified into different topics with reference to negative sentiment words. The findings help the researchers to explore several issues, such as the characteristics of negative sentiment words on each topic. One contribution of this work is to build a small data collection of negative comments and sentimental words. These words provide the first step towards constructing sentimental thesauri based on Plutchik’s circumplex model (Plutchik, 1997). Another contribution is that this study initially explores several negative document classification techniques and discusses the principles and bases for classifying negative documents. The analysis enhances the understanding of the ways in which people present negative sentiments in social networks.

**Literature review**

**Social media topic classification and semantic text analysis**

In recent years, some studies have analysed tweet topic categorisations. For example, Honeycutt and Herring (2009) used a grounded theory approach on their sample and found 12 distinct categories of tweets: about the addressee, to announce or advertise, to exhort, information for others, information for self, meta-commentary, media use, opinions, other’s experience, self-experience, to solicit information and others. Another example, Hasler et al., (2014), took an inductive or grounded approach and sorted the tweet content into 21 different topics: health condition, relationships, pregnancy, health resources, legal issues, abuse, sex, substance use, grief/death, bullying/harassment, sexuality, money issues, parenting, religions, dreams, employment, feeling lost/lonely, alternative lifestyle, caring for others, homelessness and schoolwork. In addition, the *UK Media Codebook*, developed by Jennings and Bevan (2010), classified social media topics into macroeconomics, civil right/minority issues/civil liberty, health, agriculture, labour and employment, education, environment, energy, transportation, law/crime/family issues, social welfare, community development, planning and housing issues, finance and domestic commerce, defence, space/science/technology/communications, foreign trade, international affairs/foreign aid, government operations, public lands/water management/colonial and territorial issues, regional and local government administration, weather/natural disasters, fires/accidents/other manmade disasters, arts/history/culture/entertainment, sports/recreation, deaths/death notices and obituaries, churches/religions, political parties, human interests, news in brief, picture gallery, display advertising, etc. These categories are various and distinct. However, there exist few research studies on categorizing negative tweets.

In addition, several scholars have worked on building the ontology of microblogs, which will assist in making microblog classifications more effective. They believed that Web
ontology or semantic networks could be used to conduct semantic text analysis (Cambria et al., 2013; Grassi et al., 2011; Olsher, 2012) and help in comprehending the conceptual and effective information associated with natural language opinions. The concept-based approach relies heavily on the depth and breadth of the knowledge ontology. Without a comprehensive resource encompassing human knowledge, an opinion mining system will have difficulty grasping the semantics of natural language text. Abel et al., (2011) introduced approaches for enriching the semantics of Twitter posts and modeling users based on their microblogging activities. Magumba and Nabende (2017) built an ontology of disease-related concepts designated for detection of disease incidence in tweets. This study provides a base for scholars to further conduct sentiment analysis on social media data, such as tweets.

*Semantic sentiment analysis on tweets*

Twitter is a microblogging site on which users can post updates (tweets) to friends (followers). It is an efficient tool for people to express their opinions or emotions, such as movie or general product reviews (Blitzer et al., 2007; Dave et al., 2003; Pang et al., 2002), political commentary (Lin et al., 2006; Thomas et al., 2006), news comments (Devitt and Ahmad, 2007), etc. It has become an immense data set for sentiment analysis. Sentiment analysis, though it has already become a popular research topic in recent years, remains a relatively new research field. Therefore, there is still much room for further research in this area.

There is a consensus among researchers that better results are obtained in sentiment analysis by using supervised learning techniques, such as naive Bayes, maximum entropy, latent dirichlet association (LDA), quadratic discriminate analysis (QDA) and support vector machines (SVM). For example, Barbosa and Feng (2010) found that SVMs yield better results, while others, such as Park and Paroubek (2010), advocated naive Bayes and reported good results for the maximum entropy classifier.

Except for the aforementioned machine learning techniques, some scholars use feature models to conduct twitter sentiment analysis and classification, such as the following:

1. **Bag-of-words**: The bag-of-words model is one of the most widely used feature models for almost all text classification tasks, due to its simplicity and good performance. The simplest way to incorporate this model in a classifier is by using unigrams as features (Baqapuri, 2016). Another feature vector is term presence and frequency. Term frequencies have traditionally been important in standard information retrieval. The popularity of term frequency-inverse document frequency illustrates the presence of a binary-valued feature vector in which the entries indicate only whether a term occurs (value 1) or does not (value 0). In Saif et al., (2016), a lexicon-based approach for sentiment analysis on Twitter is described. A model for describing the context of tweets is proposed. The model takes into account several aspects and their interrelations and allows for the detection of sentiment at both entity-level and tweet-level.

2. **Grammatical features**: Grammatical features analysis is used in natural language processing, as parts-of-speech is commonly used in this domain. The concept involves tagging each word of a tweet according to which part of speech it belongs: noun, pronoun, verb, adjective, adverb, interjections, intensifier, etc. In addition, some scholars contend that handling negation can be an important concern in opinion- and sentiment-related analysis. While the sentence representations of “I like riding the bicycle” and “I don’t like riding the bicycle” are considered to be very similar by most commonly used similarity measures, the only differing token,
the negation term, forces the two sentences into opposite classes. Fersini et al., (2016) investigated the impact of expressive signals, such as adjectives, pragmatic particles and expressive lengthening, on sentiment polarity classification on Twitter. Expressive signals have been used to enrich the feature space of baseline and ensemble classifiers. Experimental results show that only adjectives play a fundamental role as expressive signal.

(3) **Sentiment is dependent on context:** In previous research, researchers believed that the sentiment polarities are dependent on topics or domains. The same word may have different sentiment polarities in different domains. For instance, though the adjective “complex” in the sentence, “The movie is complex and great!” may have positive orientation in a movie review, it could also have a negative orientation, as for example in the sentence, “It is hard to use such a complex camera” in an electronics review. “Hate” represents high negative emotion in some fields, such as attitude towards people, attitude to life, or to exhort, but it is not always used in all fields for people to express bad emotions. Therefore, it is more suitable to analyse the topic or context and sentiment simultaneously. Li et al., (2010) assumed that sentiments are related to the topic in documents and put forward a joint sentiment and topic model; finally, they observed that sentiment orientation of each word is dependent on the local context.

**Sentiment words analysis**

Some research discussed what sentiment words are used within particular topics. For instance, Borth et al., (2013) performed tag frequency analysis to obtain the top 100 tags. For example, the tags of different emotions are as follows:

- joy (joy, happy, love, smile, beautiful, flowers, light, nature, kids, Christmas);
- terror (terror, horror, zombie, fear, dark, street, Halloween, war, undead, bomb);
- amazement (amazing, beautiful, nature, wonder, light, love, sky, eyes, clouds, landscape); and
- disgust (disgusting, gross, food, nasty, sick, dirty, dead, face, blood, insect).

On the other hand, some scholars classify documents into positive and negative. For example, tweet sentiment visualisation tools – known as sentiment viz – can categorize the document into pleasant and unpleasant by identifying a few sentiment words, such as pleasant: alert, excited, elated, happy, contented, serene, relaxed, and calm; and unpleasant: tense, nervous, stressed, upset, sad, unhappy, depressed, and bored. Yang (2015) focussed on life satisfaction and classified related topic tweet words into different categories, such as positive emotion \{love, nice, sweet\} and negative emotion \{hurt, ugly, nasty\}. The negative emotion categories also include anxiety \{worried, fearful\}, anger \{hate, kill, annoyed\} and sadness \{crying, grief, sad\}.

Bertola and Patti (2016) used the affective model which is based on ontology of emotion to detect emotions elicited by art works via social media analysis. A classical ontology of emotion is inspired by Plutchik’s (1997) circumplex model, a well-founded psychological model of human emotions, which includes emotions such as joy, trust, fear, surprise, sadness, disgust, anger and anticipation. Another six basic emotions proposed by Ekman (1971) are used to investigate the impact of emotions on social media author profiling in Rangel and Rosso (2016).
As sentiment analysis of tweets is a relatively new research topic, most research studies on tweet sentiment analysis focuses on comparing the efficiency of various machine learning techniques, such as LDA, SVM, naive Bayes, maximum entropy etc. or applying those analysis methods to different tweet topics and comparing the results. There is little prior research on sentimental information character analysis, especially finding the features of negative tweets on Twitter. The present findings could help to fill this gap.

**Methodology**

*Data resource*

Tweet data were harvested from Twitter using TREC Microblog Track 2013. The participants provided a corpus of approximately 240 million tweets, which were collected over a two-month period from 1 February to 31 March 2013.

*Data processing*

Data processing was conducted through four steps. First, tweets were collected through Twitter’s application programming interface. Second, negative sentiment tweets were determined by calculating how many negative sentiment words are in one tweet; a negative word list would be given, if the tweet has negative sentiment words, the weight is \( n \). Third, having identified the top 1,000 tweets which have the most sentiment words and analysed them using manual recognition, 787 valid tweets and 213 invalid tweets were derived in succession. Tweets were deemed to be invalid when their contents were meaningless or they were not in fact negative sentiment tweets. Fourth, the tweets were classified into different topics. The data processing procedures are presented in Figure 1. At this stage, the research undertook an inductive or grounded approach (Glaser and Strauss, 1967), using categories to enable identification of patterns at both general and more specific levels. In some tweets, topics are relatively easy to identify because they are referred to by name or by an associated term (e.g. “cancer”, “depression”, “muscle ache” indicates “health”) or by the prominence of language obviously associated with a topic (e.g. “I hate my work” coded as “working and studying”). Ten different categories were recognized and are summarized in Table I. The categories build on several preliminary classification studies identified by some scholars who classified tweets into several topics, such as health condition, relationship, money issues, religion, dreams, employment, school work, etc. (Bright, 2016; Hasler *et al.*, 2014).

![Figure 1. Data processing procedures](image-url)
<table>
<thead>
<tr>
<th>No.</th>
<th>Categories</th>
<th>Description</th>
<th>Tweet example</th>
</tr>
</thead>
</table>
| 1   | Politics and society       | Negative sentiment comments on countries' politics, or leaders, terrorist incidents, corruption, economics, war, and so on | [1] GOP's hackneyed economic agenda is dismaying (but at least imp. issue.) Their hateful abortion agenda is despicable, irrelevant. #dearjohn  
[2] @IMZandor yeah mum told me about this shit in Egypt, I don’t watch/listen to news but she does, sad, pathetic and sad. Fuck politics |
| 2   | Health                     | People suffering maladies, such as fevers, cold, poor sleep, and so on       | [1] What a way to end the 1st month of 2011 […] Sick again! Throat giving me lots of problem, fever and muscle aches.. Irritated!  
[2] Oh bad night sleep, got a bloody cold, sore throat, and then had a distressing dream and been up since 4am, oh dear god, bloomin gr8 weekend |
| 3   | Working and studying       | Complaints regarding difficulties or boredom associated with jobs or school work | [1] I’m still up. Ugh. /i wanna go to bed. :/ Too bad I still have all this nonsense to do. Blah, we better not have school tomorrow. #tired  
[2] I hate my work. I hate my Boss. I hate being sick […] TOO MUCH HATE in the air […] |
| 4   | Attitude to people         | Self-criticism and comments on others                                        | [1] I’m selfish, impatient and a little insecure, I make mistakes, I am out of control and at times hard to handle […]  
[2] I dunno why the hell I miss you, Cause all you do is make me ache and hurt, The pain you cause is […] |
| 5   | Attitude to life           | Expressing dislike of one’s present life and expressing negative feelings     | [1] I’m tired of my brother, I’m tired of my sister, I’m tired of this house, I’m tired of this town, I’m tired […]  
[2] #deep RT @ispeakfemale Afraid. That’s what we always are. Afraid to love, afraid to share, afraid to open up, afraid to care. That’s life |
| 6   | Routine life               | Complaints about routine, trivial affairs, such as pet-raising, housework, noisy neighbours, and so on | [1] and […] we don’t have a normal doorbell; we have a super noisy “DING-DONG DING-DANG. DING-DANG-DONG-DING” musical one.  
[2] Damn my caring for cats. You know, a sickly meow floating in through my cracked window sucks. Poor things probably cold as all hell |
| 7   | Exhort                     | Words to exhort people and explain why negative emotions emerge              | [1] “Wow@JeffDauler: Pain in life is inevitable but suffering is not. Pain is what the world does to you, suffering is what you do to yourself”  
[2] @jmalonzo: RT @NoDissassemble: Fear leads to Risk. Risk leads to Process. Process leads to Hate […] and suffering, and Gantt charts (continued) |
There are many different kinds of negative word lists. Godbole et al., (2007) constructed positive and negative common seed words, which are extended from the WordNet’s synonyms and antonyms. Neviarouskaya et al., (2011) formulated the emotional word set SentiFul, which also extracted from and expanded WordNet’s English word ontology. Heerschop et al., (2011) compared the effects of three kinds of emotion word sets all based on WordNet: the first is an expansion of the WordNet emotional vocabulary using seed words, the second extracts an emotional word set from WordNet through the PageRank semantic algorithm and the third is the existing emotional word set SentiWordNet. The experimental results showed that SentiWordNet was more effective in classifying sentiment than the other two sets. Mohammad et al., (2009) extended the emotional word set by adding English word prefixes and suffixes. After comparing several negative word lists, Hu and Liu’s (2004) negative words set was chosen for this research. These negative words (n = 4,781) were previously used in research analysis of social network context sentences, especially on Twitter.

**Coding reliability**

This paper used bottom-up thinking to establish the negative tweets classification architecture. To ensure the validity of the coding scheme, another coder, who is an information science master’s student, was invited to perform secondary coding. Intercoder reliability is calculated using the following formula:
\[ A = M / \left( \left( \sum_{i=1}^{n} N_{i} \right) / n \right) \]

where \( A \) is agreement; \( M \) is the number of coding events agreed by all coders; \( N_{i} \) is the number of coding events assigned by \( i \)th coder; and \( n \) is the number of coders (Kracker and Wang, 2002). After performing this calculation, an internal consistency of coding was 91.6 per cent.

**Findings**

Taking a grounded approach, the tweets were divided into ten topics based on the contents from the 787 valid tweets: “politics and society”, “health”, “working and studying”, “attitude to people”, “attitude to life”, “routine life”, “exhort”, “entertainment and sports”, “traffic” and “others”. The findings refer to other scholars’ classifications (Bright, 2016; Hasler et al., 2014). Description and examples of the negative tweet topic categories are presented in Table I. The percentage of all these topics, except the topic “others”, are shown in Figure 2.

**Categories’ description**

1. **Politics and society**: The content of these tweets include, for example, comments on a country’s leader; terrorist incidents in Egypt, Pakistan and Iraq; incidents of corruption; violent crimes; and suicide. The study identified two kinds of words: “related words” and “negative words”. Related words represent the facts on which users comment, which are more related to the context of the tweets. Negative words are normally used to express negative emotion. The related and negative words are listed in Table I:
   - **Related words**: Nigga; corruption; dictator; Egyptian; Irish; Moscow; potatoes; racism; rhetoric; war; yard; agenda; animals; Australian; bank; black; bleed; bomber; break; and burn.
   - **Negative words**: The research found some negative words, such as shit, fuck, bitch, etc., that are basically meaningless so these are gotten rid of and the more useful negative words are kept suicide; killed; death; poor; attack; lie; sad; violent; bombing; chaos; criminal; death; dictator; expensive; fall; famine; ignorant; poisoning; protest; and starve.

2. **Health**: It includes the feeling from some people suffering specific diseases, bad sleep, descriptions of symptoms of a disease, etc.
(3) Working and studying: This mainly describes the human being’s complaints with difficulties or troubles of working and studying.

(4) Attitude to people: It mainly expresses the two kinds of mood contributing towards attitudes for people, one is self-criticism and the other is negative comments on other people.

(5) Attitude to life: The content about these tweets includes people’s complaints about life.

(6) Routine life: It includes housework, pets-raising, traveling, etc.

(7) Exhort: It explains why those negative emotions emerged and some tweets give suggestions to overcome those bad emotions, such as fear, angry, stress, etc.

(8) Entertainment and sports: The content about these tweets include bad comments on entertainment, such as movies, TV shows, some negative emotions on sports games, etc.
- **Related words**: Stadium; fan; England; batsman; Kewell; JYJ; bear; fans; footballer; player; ball; Drogba; scored; Ronaldo; football; chill; marathon; lost; noise; and match.

- **Negative words**: Hard; loser; miserable; disappointed; shocked; cruel; hoax; break; defense; error; danger; hate; hell; fool; fake; failure; die; and scared.

(9) **Traffic**

- **Related words**: Jam; kmps; delay; bus stop; joint; cold; lie; drunk driver; slowly; surfer; driving; traffic; and hyperbole.

- **Negative words**: Die; anoying; shit; lie; fucking; stupid; dumb; fuck; jam; fucking; sick; silly; idiots; jam; nonsense; and blah.

**Part of speech analysis**

The distribution of the part of speech of negative sentimental words is computed and reported in Table II and Figure 3. From the view of the total number, the adjective is considered to be very important, with a high probability, which is consistent with Fersini et al.’s (2016) conclusion. That is, people usually use adjectives to express their negative affection. In some specific fields, there are different quantities. The quantity of interjection or slang is more than adjectives in the category of routine life. In the

<table>
<thead>
<tr>
<th>No.</th>
<th>Category</th>
<th>Verb</th>
<th>Noun</th>
<th>Interjection or slang</th>
<th>Adverb</th>
<th>Adjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Politics and society</td>
<td>48</td>
<td>50</td>
<td>24</td>
<td>3</td>
<td>64</td>
</tr>
<tr>
<td>2</td>
<td>Health</td>
<td>16</td>
<td>24</td>
<td>11</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Working and studying</td>
<td>34</td>
<td>14</td>
<td>13</td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>Attitude to people</td>
<td>33</td>
<td>35</td>
<td>28</td>
<td>2</td>
<td>56</td>
</tr>
<tr>
<td>5</td>
<td>Attitude to life</td>
<td>15</td>
<td>7</td>
<td>5</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Routine life</td>
<td>30</td>
<td>15</td>
<td>76</td>
<td>14</td>
<td>74</td>
</tr>
<tr>
<td>7</td>
<td>Exhort</td>
<td>21</td>
<td>36</td>
<td>1</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Entertainment and sports</td>
<td>23</td>
<td>12</td>
<td>8</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Traffic</td>
<td>9</td>
<td>23</td>
<td>5</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Others</td>
<td>118</td>
<td>252</td>
<td>317</td>
<td>7</td>
<td>237</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td>352</td>
<td>475</td>
<td>488</td>
<td>26</td>
<td>624</td>
</tr>
</tbody>
</table>

Table II. Part of speech analysis

**Figure 3.** Part of speech distribution
categories of politics and society, as well as entertainment and sports, people prefer to use the verb of a negative sentimental word to express their negative emotion. Negative nouns are widely used in the fields of politics and society, attitude to people, exhort and traffic.

Density of negative words

The use of negative words could reflect the degree to which the user is releasing negative emotions. To explore and measure the strength of the negative emotion, the study introduces an indicator named density of negative words. Density can be calculated by the ratio of the number of negative words to the total number of words in each tweet. The formula is as follows:

\[ D = \frac{N}{T} \]

where \( D \) is density of negative words, \( N \) is the number of negative words and \( T \) is the total number of words. The density of negative words for all 787 tweets is 0.262, which reflects the average density of negative words.

The average density can be compared with the density of negative words for each topic. If the density for one topic is higher than 0.262, it means that users express stronger negative emotion in relation to this topic; it also means that negative emotions are concentrated within this topic. The negative word densities for each of the different topics are shown in Table III. Figure 4 shows that the densities of routine life, entertainment and

<table>
<thead>
<tr>
<th>Category</th>
<th>T</th>
<th>N</th>
<th>D</th>
<th>D-0.262</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics and society</td>
<td>1,686</td>
<td>407</td>
<td>0.241</td>
<td>-0.021</td>
</tr>
<tr>
<td>Health</td>
<td>970</td>
<td>243</td>
<td>0.251</td>
<td>-0.01</td>
</tr>
<tr>
<td>Working and studying</td>
<td>443</td>
<td>119</td>
<td>0.269</td>
<td>0.007</td>
</tr>
<tr>
<td>Attitude to people</td>
<td>988</td>
<td>268</td>
<td>0.271</td>
<td>0.009</td>
</tr>
<tr>
<td>Attitude to life</td>
<td>351</td>
<td>82</td>
<td>0.234</td>
<td>-0.028</td>
</tr>
<tr>
<td>Routine life</td>
<td>1,333</td>
<td>374</td>
<td>0.281</td>
<td>0.019</td>
</tr>
<tr>
<td>Exhort</td>
<td>577</td>
<td>151</td>
<td>0.261</td>
<td>-0.01</td>
</tr>
<tr>
<td>Entertainment and sports</td>
<td>365</td>
<td>106</td>
<td>0.290</td>
<td>0.028</td>
</tr>
<tr>
<td>Traffic</td>
<td>217</td>
<td>72</td>
<td>0.332</td>
<td>0.07</td>
</tr>
<tr>
<td>Total tweets</td>
<td>6,941</td>
<td>1,822</td>
<td>0.261</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table III. The density of negative words

Figure 4. Negative word density for each topic plotted against the average density
sports and traffic are higher than the average density 0.261, which means people are more compliant about these areas.

Table III and Figure 4 show that negative word densities for some topics are higher than the average density, such as working and studying, routine life, attitude to people, entertainment and sports and traffic. On these topics, users tend to use negative words to directly express negative emotions concerning people and events. Furthermore, these topics are the hot spot through which people release their negative sentiment. However, the negative word densities for some topics are lower than average, such as politics and society, health and exhort. One of the potential reasons is that people use more neutral words to describe facts in these areas rather than expressing negative emotion by using negative words.

The high-frequency negative sentiment words

The frequency of negative sentiment words was counted in the 787 sample tweets. The top 50 words are represented in Table IV. “Hate” was found most frequently, appearing 104 times in the 787 tweets.

Negative word distribution

Table IV clearly shows that some slang words are always used in negative tweets, such as “shit”, “fuck”, “fucking”, “damn” and so on. However, there are many other meaningful words that are valuable and should arouse people’s attention; for example, “hate”, “bad”, “hard”, “jam”, “lie”, “pain”, “death”, “lost”, “sick”, “suicide” and so on. Among those negative words, some appear with quite high frequency in the entire 787 tweets’ sample. The visualisation tool AntCount was used to present the word distribution in tweets (Figure 5). The distribution of the negative words shows the frequency of each word’s use in relation to each of the ten topics. From left to right, the ten topics are ordered as politics and society, health, working and studying, attitude to people, attitude to life, routine life, exhort, entertainment and sports, traffic and others.

Furthermore, the findings revealed that some negative words are routinely used in relation to particular topics (Figure 6). For example, most people use “jam” in the traffic topic, “suicide” is frequently used in attitude to people and attitude to life topics, and “death” is often mentioned in the health category.

Distribution of negative words obeys Zipf’s law

According to Zipf’s law, the most frequent word in a corpus occurs twice as often as the second most frequent word in the corpus, three times as often as the third-most frequent word, etc. In a histogram sorted by word rank, with the most frequent words first, the shape of the curve is a Zipf curve (Manning and Schuetze, 1999). Zipf’s law states that if the Zipf curve is plotted on a log-log scale, it must be a straight line with a slope of $-1$. The study’s results demonstrate that the distribution of negative words in the sample data obey Zipf’s law (Figures 7 and 8).

Discussion

Several exciting directions for investigation are presented by the research findings. People use different related words to describe an incident or circumstance. These words are hot spots on which Twitter users always focus in these areas during a given period. The negative words of each category help people to release emotions, such as anger and depression. For example, in health, people use “attack”, “hurt”, “depression”, “sad” and
“terrible” to express their feelings effectively; in relation to working and studying, they use “hate”, “suck”, “bad” and “crazy” to strike a responsive chord with other people; in the attitude to life category, some people convey their mood without reservation using “disaster”, “ugly”, “stupid”, etc., whereas others exercise greater control over their
sentiments by using more reserved words, such as “disappointing” or “boring”. Some people like using multiple instances of “fuck”, “fucking” or “hate” to strongly express dissatisfaction. Hence, different people’s personalities should be considered when analysing negative sentiment tweets, as different personality types tend to convey how they feel in different ways. Furthermore, the part of speech of negative words is different in each topic. The experiment resulting in Table II shows that adjectives are commonly used in negative sentimental tweets, which is consistent with Fersini et al.’s (2016) conclusion. But in some areas, noun, verb, interjection or slang are widely used. The quantities of those words are even more than adjectives. Table III shows that the utilisation rates of negative words in routine life, entertainment and sports and traffic are obviously higher than those for other fields. Meanwhile, the research finds that negative words’ distribution varies between Figure 4 and Figure 5. Some negative words are used in some specific areas, not in all fields. It can be seen that it is necessary to analyse the characteristics of negative tweets from different fields.

![Table II: The distribution of the negative word in 787 tweets](image1)

**Figure 5.** A sample of the negative words most routinely used in negative tweets

![Table III: The distribution of the negative word in 787 tweets](image2)

**Figure 6.** Some negative words are always used in some fixed topics

![Zipf curve](image3)

**Figure 7.** Zipf curve
This research provides a preliminary foundation to detect people’s emotion by related sentimental words. Further research will construct sentimental thesauri, which helps to understand a person’s affectation in Twitter by identifying sentimental words, judging whether the tweet is negative or positive. The sentimental thesauri could help recognize the tweets’ affectation by machine learning techniques, such as the SVM model. Meanwhile, this work notes that not all the sentimental words appear homogeneously in all fields. In the previous findings, some negative words are used more frequently in certain fields. Therefore, it is necessary to think about the context and build local and global sentimental thesauri which make sentiment analysis more effective. The concept of a local sentimental thesauri borrows from the idea of local analysis in Xu and Croft’s (1996) paper. They have compared the methods of local analysis and global analysis in query expansion. Local analysis, with the use of context and phrase structure, is generally more effective than global analysis. In this research, some negative words, such as “hate”, “shit”, “fuck”, “bad”, “pain”, “lost” and “stupid”, are widely used in all fields. The study refers to them as a global sentimental thesauri, however, local sentimental thesauri are always used in some specific areas, such as “sick”, “bad”, “die”, and so forth are always used in the health field and “lie”, “suicide”, and so on are used in the attitude to life topic. Those local sentimental thesauri could improve the accuracy of text classification. The findings of the paper are consistent with the findings of the research of Li et al. (2010). It is more suitable to analyse the topic and sentiment simultaneously. People need to analyse the sentiment in the more detailed topic or domain level.

There are series of theories about semantic analysis that have been proposed in the past. Relatively speaking, very limited research studies do sentimental analysis on tweets. This work is an effort towards this direction. Ideally, Plutchik’s (1997) circumplex model could be expanded to apply to sentimental analysis. This study may provide the first stepping stone towards this direction. In future research, the researchers could combine with Plutchik’s model and analyse people’s feelings in more detail not only in the negative sentiment field but also for some specific emotions, such as joy, trust, fear, surprise, sadness, etc. Further study will be conducted in the future.

Conclusion
In this paper, many categories of negative sentiment tweets in Twitter have been explored. The related words which have been identified could be used to construct a framework of negative sentiment tweets. Further, the paper also revealed several features of negative tweets, especially in relation to each different topic, exploring what kinds of negative words
people use and why they choose those words to express their negative emotions. Additionally, other features of negative sentiment tweets were found in the research, such as the part of speech, density and frequency and distribution of negative words. The small negative words set formulated in this study could be used by other researchers to conduct further tweet sentiment analysis. Another important contribution of this paper is that sentiment analysis method is not only suitable for negative tweets but also for positive and neural tweets; more detailed classification of emotions will help future research into information retrieval and organisation.

However, the research has some limitations. A small sample of 787 negative tweets was analysed manually in this paper. Therefore, a large data sample should be considered in the future. In further research, tweets can be classified through a machine learning method by inputting the feature words described in this research.

References


**About the authors**

Ling Zhang is an Associate Professor at the School of Management, Wuhan University of Science and Technology. Dr Zhang’s research interests include knowledge management and social network.

Wei Dong is an Associate Professor at the School of Education, Tianjin University. He has conducted research on knowledge management and usability of websites. Wei Dong is the corresponding author and can be contacted at: weixiong83@163.com

Xiangming Mu is an Associate Professor at the School of Information Studies, University of Wisconsin-Milwaukee. Dr Mu’s research interests include user behaviour study and information retrieval.

---

For instructions on how to order reprints of this article, please visit our website: [www.emeraldgrouppublishing.com/licensing/reprints.htm](http://www.emeraldgrouppublishing.com/licensing/reprints.htm)

Or contact us for further details: permissions@emeraldinsight.com