Jurek *et al. Secur Inform (2015) 4:9*
DOI 10.1186/s13388-015-0024-x

Security Informatics
a SpringerOpen Journal

RESEARCH Open Access

Improved lexicon-based sentiment analysis for social media analytics

Anna Jurek*, Maurice D. Mulvenna and Yaxin Bi

Abstract

Social media channels, such as Facebook or Twitter, allow for people to express their views and opinions about any public topics. Public sentiment related to future events, such as demonstrations or parades, indicate public attitude and therefore may be applied while trying to estimate the level of disruption and disorder during such events. Consequently, sentiment analysis of social media content may be of interest for different organisations, especially in security and law enforcement sectors. This paper presents a new lexicon-based sentiment analysis algorithm that has been designed with the main focus on real time Twitter content analysis. The algorithm consists of two key components, namely sentiment normalisation and evidence-based combination function, which have been used in order to estimate the intensity of the sentiment rather than positive/negative label and to support the mixed sentiment classification process. Finally, we illustrate a case study examining the relation between negative sentiment of twitter posts related to English Defence League and the level of disorder during the organisation's related events.

Keywords: Sentiment analysis, Social media, Security

Background

Social media is one of the most significant information exchange technology of the 21st century. People of all ages use social media to post messages, photos and videos about their daily activities. Social media channels, such as Twitter and Facebook, provide very convenient and efficient ways of communicating and sharing information publically. Consequently, the role of social media in crime investigation and prevention is growing rapidly. Social media are rapidly becoming a source of information for early warning systems in public safety. According to the LexisNexis report [1] four out of five law enforcement professionals utilise social media for investigation purposes. According to the statistics given in [1], 69 % are using social media tools for gathering information about crimes and about 41 per cent are using social media for crime anticipation.

Sentiment analysis has been already applied in a number of different, non-security domains for monitoring and forecasting public opinions. In [2] the authors applied a domain-specific lexicon in order to classify customer reviews of hotels into five star categories. Sentiment analysis performed on Twitter was applied in [3] in an effort to forecast box-office revenues for movies. Following their study it was found that there was a relationship between the rate of movie tweets and the real-world box-office performance. A similar application of social media analysis was presented in [4]. In their study the authors uncovered a relationship between online discussion on the Internet Movie Database and the Academy Awards nominations and the box-office success. In [5] the authors developed a new model for event analytics. Their proposed framework characterizes the segments and topics of an event via Twitter sentiment. In their study they focused on two public events, namely the U.S. Presidential debate in 2012 and President Obama's Middle East speech in 2011. The application of sentiment analysis in the tourism domain was presented in [6]. The authors introduced the use of lexicon databases for sentiment analysis of user reviews acquired from TripAdvisor for accommodation and food. In [7] social media was presented as a new opportunity to study bullying in the physical and cyber worlds. In their study the authors developed a text classification model that recognised

*Correspondence: jurek-a@email.ulster.ac.uk
Faculty of Computing and Engineering, School of Computing and Mathematics, Ulster University, Newtownabbey BT37 0QB, UK

different emotions (anger, embarrassment, empathy, fear, pride, relief, sadness) in Twitter posts. In [8] Twitter data was applied in an effort to identify the correlation between public and market sentiment. The authors classified messages into four different mood classes, namely *calm*, *happy*, *alert* and *kind*. Following this, the identified moods and previous day's Dow Jones Industrial Average were used to predict future stock movements.

The aforementioned studies demonstrate that social media and sentiment analysis have been considered in many different application domains. Some research has been already directed towards designing social media-based intelligent systems for the purpose of supporting decisions in the area of public safety. In [9] a topic detection technique was proposed that allows the retrieval in real-time of the most emerging topics expressed by a community through social media. A probabilistic model was developed in [10] that can predict the risk of falling ill for individuals on the basis of their social ties and co-location with other people, as revealed by their Twitter posts. Twitter corpus was also suggested as a source of information that can be applied in monitoring the diffusion of an epidemic disease such as seasonal influenza [11]. The general problem of Web-based security informatics was addressed in [12]. In their work the authors referred to three fundamental objectives, namely the discovery of security-relevant data and information, target situational awareness and predictive analysis. They proposed an analysis methodology and evaluated it through a series of real-world examples, such as detection of cyber incidents in near real-time, estimation of public opinions in contentious situations, discovery of emerging topics and trends, and early warning analysis for mobilization and protest events. In [13] Twitter data was applied in order to detect online communities involved in conversations around the 2013 Syrian Sarin gas attack topic. Following this, different types of leaders were identified within the communities. A work related to predicting popularity of forum threads related to public events was undertaken in [14]. In [15] a method based on trigger keywords and contextual cues was proposed for detecting threatening messages on social media. A Violence Detection Model was proposed in [16] for identification of violence related topics being discussed on a micro blog. Social media traffic around the Great Eastern Japan Earthquake was analysed in [17] in order to investigate the relation between people's activities and the series of events occurring in the event's aftermath. Similar research with respect to the 2011 Tohoku Earthquake has been undertaken in [18], where both, English and Japanese tweets were analysed to determine different reaction attitudes between local and foreign residents.

To date, little research has focused on inferring the sentiment of social media content for the purpose of security analysis. In [19, 20] lexicon-based sentiment analysis algorithms were introduced and presented in a number of different case studies. In [19] the potential of the methods was illustrated by estimating regional public opinion regarding two events: the 2009 Jakarta hotel bombing and the 2011 Egyptian revolution. In [20] authors investigated the relationship between regional online sentiment about Palestinian suicide bombing attacks against Israel and actual bombing events. In the same work they also studied the impact of public sentiment on the epidemic risk of H1N1 vaccination. Sentiment analysis was applied in [21] for identifying top malware sellers and stolen credit card sellers in the online underground economy. Public opinion around the 2012 Pussy Riot event was evaluated through sentiment analysis of social media posts in [22]. In some work [23, 24], sentiment analysis was studied as a technique for detecting radicalisation in social media. In work presented in [23], sentiment analysis together with lexical and social network analysis was applied to examine and characterise the users of radicalised forums. In [23] sentiment analysis was suggested as one of a set of linguistic markers that could be applied for identifying potential lone wolf terrorism.

In this work we focus on the application of sentiment analysis of Twitter content in estimating the level of disruption and disorder during public events. We developed a lexicon-based sentiment analysis algorithm that differs from existing models in the way that it aggregates the sentiment values of positive and negative words within a message. Through the application of a normalisation function the sentiment of a message is represented as a value from a range of −100 to 100. Consequently, a more comprehensive analysis can be undertaken regarding the sentiment as opposed to positive–negative-neutral classification. Besides this, such an approach is more appropriate for real time analysis given that it allows detailed visualisation of the sentiment over time. In an effort to increase the accuracy of the algorithm we proposed an evidence-based combination function that is applied in the case when positive and negative words co-occur in a message. Furthermore, a modified manner of handling negation and intensification within a message was introduced. Finally, we illustrate a case study examining the relationship between sentiment about English Defence League (EDL) prior to EDL demonstration and the level of disruption and disorder during the event. The method has been already introduced in one of our previous papers [32]. In this manuscript we described in much more detail the theoretical aspects of the approach. We explained step by step how the sentiment normalisation function had been developed. Following this, we

evaluated how the method performed with long messages, such as movie reviews. In this work, the algorithm has been modified in the way that it can perform sentiment analysis on sentence level. In this manner we wished to investigate how sentence level analysis affects the sentiment's classification accuracy.

The paper is organised as follows. In the following section we present the new lexicon-based approach. Results of the empirical evaluation of the algorithm are demonstrated in "Empirical evaluation". In "Discussion" we illustrate the case study followed by summary and future work in "Case study: English defence league".

Lexicon-based sentiment analysis

Application of a lexicon is one of the two main approaches to sentiment analysis and it involves calculating the sentiment from the semantic orientation of word or phrases that occur in a text [25]. With this approach a dictionary of positive and negative words is required, with a positive or negative sentiment value assigned to each of the words. Different approaches to creating dictionaries have been proposed, including manual [26] and automatic [27] approaches. Generally speaking, in lexicon-based approaches a piece of text message is represented as a bag of words. Following this representation of the message, sentiment values from the dictionary are assigned to all positive and negative words or phrases within the message. A combining function, such as sum or average, is applied in order to make the final prediction regarding the overall sentiment for the message. Apart from a sentiment value, the aspect of the local context of a word is usually taken into consideration, such as negation or intensification.

In our work we have decided to apply a lexicon-based approach in order to avoid the need to generate a labelled training set. The main disadvantage of machine learning models is their reliance on labelled data. It is extremely difficult to ensure that sufficient and correctly labelled data can be obtained. Besides this, the fact that a lexicon-based approach can be more easily understood and modified by a human is considered a significant advantage for our work. We found it easier to generate an appropriate lexicon than collect and label relevant corpus. Given that the data pulled from social media are created by users from all over the globe, there is a limitation if the algorithm can only handle English language. Consequently, sentiment analysis algorithm should be more easily transformable into different languages. Later in the paper we discuss how a lexicon-based sentiment analysis algorithm can be adapted to different languages by an appropriate translation of the sentiment lexicon and application of string similarity functions.

The following five sub-sections describe in details the development of the algorithm applied in this study.

Sentiment lexicon

The sentiment lexicon constructed contains about 6300 words. It was generated manually with application of SentiWordNet [28] as a baseline. Each word in the lexicon has assigned a value representing sentiment in the range of −100 (most negative) to 100 (most positive). From an empirical knowledge it is known that some of the positive and negative words sometime occur with neutral meaning in a sentence context. For example, sentence "Enjoying my lazy Sunday!!" represents a positive message that contains one positive (enjoying) and one negative (lazy) word. It may be difficult in such a case to decide between positive and negative. In an effort to alleviate this issue, besides the sentiment value, for each word from the lexicon we estimated a conditional probability (denoted by P) as presented in Eq. 1.

$$\begin{aligned} &P(positive \mid w) \; for \; positive \; w \\ &P\big(negative \mid w\big) \; for \; negative \; w \end{aligned} \tag{1}$$

Based on a set of labelled data, for each positive word we estimated the probability that a random message containing this word is positive. In the same manner the probabilities were estimated for each negative word. We intended to investigate if application of such information in the sentiment classification process can help to handle messages with mixed (positive and negative) sentiment. For the purpose of calculating the probabilities we applied a training data set provided by Stanford [29] that contains 1.6 million (including 800,000 positive and 800,000 negative) labelled tweets. The training dataset was created automatically based on the absence of emoticons within a message. It was assumed that any tweets with positive emoticons were positive and tweets with negative emoticons were negative. List of emoticons was applied as query for Twitter API and the collected messages were automatically labelled as positive or negative, depending on the type of emoticon they contained. The process of calculating the probabilities has been performed as follow. A sample of 100,000 positive and 100,000 negative tweets has been selected randomly. Following this, for each word form the lexicon, denoted as w, its frequency among the selected positive and negative messages was calculated. Depending if the word was positive or negative, the conditional probability was calculated as presented in Eq. 2.

$$\begin{aligned} P\left(positive|w\right) &= \frac{P\big(positive \bigcap w\big)}{P(w)} = \frac{\#w_P}{\#w} \\ P\left(negative|w\right) &= \frac{P\big(negative \bigcap w\big)}{P(w)} = \frac{\#w_N}{\#w} \end{aligned} \tag{2}$$

where $\#w_P$ and $\#w_N$ stand for the number of messages from the sample that contains word w and are positive and negative, respectively. The two formulas were applied in order to estimate the probabilities for positive and negative words, respectively. In order to obtain more precise result, the process was repeated 100 times and the average probability obtained for each word has been stored in the lexicon. The probabilities are referred to as pieces of evidence later in the paper.

Negation

The most common approach to handling negation with a lexicon-based approach is by reversing the polarity of the lexicon item that stands next to the negator in a sentence [30] (e.g. good: 100 and not good: −100). In our work we proposed to take a different approach. Rather than reversing the sentiment value we proposed to formulate a negating function that calculates the sentiment value of a negated word. First, we manually created a lexicon composed of 38 negating words. Following this, applying the Twitter corpus, we selected most commonly used phrases containing negation of verbs and adjectives. In the next step, a group of 20 people was asked to rank the expressions from both of the list from most positive to most negative. Taking under consideration all the results, the final two rankings were estimated. Based on the two rankings we determined the most corresponding negating function represented as follow:

$$F_N(S) = \begin{cases} max\left\{\frac{S+100}{2},\ 10\right\} & if\ S < 0 \\ min\left\{\frac{S-100}{2},\ -10\right\} & if\ S > 0 \end{cases} \quad (3)$$

where final negation is denoted by F_N and S represents a sentiment value from the lexicon. Once a negation is recognised in a sentence, the first non-neutral word that occurs within the following three positions after the negator is searched. If a positive or negative word is identified, its new sentiment value is calculated by using Eq. (3) (e.g. enjoy: 20, do not enjoy: −40).

The advantage of our approach, in comparison to the polarity reversion, is the resulting more accurate manner of assigning the sentiment values to negated words. For instance, in sentence "I don't hate this city", the sentiment assigned to the sentence according to the inversion rule will be 100 ("hate" has value −100 in the lexicon) and the sentence will be considered as positive. In fact, it will have the same sentiment as sentence "I love this city", what is not the expected result. With the introduction of a negating function the sentiment of the sentence will be 10. As we will see later in the paper, a sentence is classified as positive if the total sentiment is greater than 25. Consequently the above sentence would not be considered as positive. Dividing the value by 2 in Eq. 2, it ensures that a very high or low sentiment cannot be obtained by negation.

Intensity

Intensifiers refer to words such as *very*, *quite*, *most,* etc. These are the words that change sentiment of the neighbouring non-neutral terms. They can be divided into two categories [29], namely amplifiers (*very*, *most*) and downtoners (*slightly*) that increase and decrease the intensity of sentiment, respectively. In our approach 25 most frequently applied intensifiers were selected and then, depending on their polarity, they were divided into 3 categories, namely downtoners, weak amplifiers and strong amplifiers. Empirically downtoners represent intensifiers that decrease value of the sentiment by 50 %. Weak and strong amplifiers increase sentiment by 50 and 100 %, respectively.

None of the negators and intensifiers is included in the sentiment lexicon. Consequently, if they appear in a sentence surrounded by only neutral text, they are considered as neutral words. However, if they appear in a neighbourhood of positive or negative words they are considered as non-neutral given that they influence the final sentiment of a sentence.

Combining function

Once all positive and negative words are identified in a sentence and their local context is verified, a combining process is performed in order to obtain the final sentiment value. In most of the existing approaches to sentiment analysis, the output of the process is represented as a positive or negative class label. In our work we attempted to design a sentiment combining function that, based on the sentiment of single words, provides the absolute sentiment of the message as a normalised value from the range of −100 to 100. The motivation for such an approach was the possibility to analyse the sentiment in the degree of intensity as opposed to positive and negative only. Apart from the polarity, we wanted to be able to determine how strongly positive/negative a sentence is and which of any two sentences is more positive/negative than the other. Consequently, the combining function should be able to model the relation between sentences depending on the number of non-neutral words and the value of the sentiment they contain. In the first attempt an average was considered as a combining function for the sentiment within a message. This solution, however, did not provide an accurate differentiation between sentences. For example, for the two sentences presented below, based on the average we are not able to recognise correctly which sentence express more positive opinion.

"The hotel is beautiful (100)."

"The hotel is beautiful (100) and the staff are outstanding (80)"

The numbers in brackets represent the sentiment values taken from the sentiment lexicon. We can find out from the above example that both words, *beautiful* and *outstanding*, are positive with sentiment values of 100 and 80, respectively. Intuitively we can say that the second sentence expresses stronger positive opinion than the first one. However, taking under consideration the average sentiment, the first sentence is more positive. Consequently, we can infer from this that the average cannot be an optimal sentiment combining function. In this study, we proposed a new normalisation formula that combines the average sentiment of a sentence and the number of words to calculate the average. The idea was that, for a given average sentiment of a message, the difference between the overall positive and overall negative sentiments should also depend on the number of positive and negative words in the message. Therefore, the overall positive/negative sentiment should be represented as a product of the average sentiment and a coefficient that's value depends on the number of positive/negative words. Following this rationale, we developed the normalisation formulas, denoted by F_P and F_N that calculate the overall positive and negative sentiment in a sentence as follow:

$$F_P = min\left\{\frac{A_P}{2 - log(p \times W_P)}, 100\right\}$$
$$F_N = max\left\{\frac{A_N}{2 - log(p \times W_N)}, -100\right\} \quad (4)$$

where A_P, A_N stand for an average of positive and negative sentiment respectively, and W_P, W_N represent the number of positive and negative words applied while calculating A_P and A_N, respectively. The idea was to apply the logarithmic function in order to model the relation between the number of positive/negative words and F_P/F_N for a given value of the average positive/negative sentiment in a sentence. The parameter p determines shape of the logarithmic function. The greater the value of p the faster the value of F_P/F_N increases as the number of non-neutral words changes. In order to determine the optimal value of p we performed a simple statistical analysis of 13,500 tweets and analyse non-neutral words' distribution across messages. Figure 1 demonstrated the results we obtained.

It can be observed that very small number of messages contain more than 3 non-neutral words. Consequently we assumed that for the value of average sentiment equals 100, the F_P/F_N is equals 100 for the number of non-neutral words being equal to 3. In order to achieve this, we need to select value p that will give a value of the coefficient in Eq. 5 equalling 1 for $W_p = 3$. The graphs in Fig. 2 demonstrate how the value of the coefficient changes for different values of p.

$$\frac{1}{2 - log(p \times W_P)} \quad (5)$$

We can observe for the table that for three non-neutral words in a message, the coefficient is equal to 1 if $p = 3.5$. Consequently, we selected $p = 3.5$ to be applied in the sentiment combining formula Eq. 4. Figure 3 demonstrates how, for an average value of sentiment being equal to 20, 40, 60, 80 and 100, the value of F_P/F_N changes for different number of non-neutral words (1…5). Each line represents value of F_P/F_N for different values of the average sentiment. We can observe, for example, that for the messages with average sentiment equalling 20, the value of F_P/F_N changes from around 15 up to 30. For the average sentiment 100, the value changes from 70 to 100.

Following the aforementioned evaluation, the formulas for calculating the overall positive and negative sentiment of a sentence were written as Eq. 6.

$$F_P = min\left\{\frac{A_P}{2 - log(3.5 \times W_P + I_P)}, 100\right\}$$
$$F_N = max\left\{\frac{A_N}{2 - log(3.5 \times W_N + I_N)}, -100\right\} \quad (6)$$

where I_P and I_N stand for the number of intensifiers that refer respectively to positive and negative words in a sentence. Instead of decreasing or increasing values of word's sentiment by 50 or 100 %, we simply decrease or increase the number of words by appropriate values of 0.5 or 1, respectively.

As an output of the sentiment combination and normalization process we obtain two values. One is from range 0–100 representing total positive sentiment of a tweet and another from range −100 to 0 standing for the total negative sentiment. Initially the algorithm compared the absolute values of the two sentiments and classified tweet as positive or negative, depending on which of the values was greater. The normalised value representing the

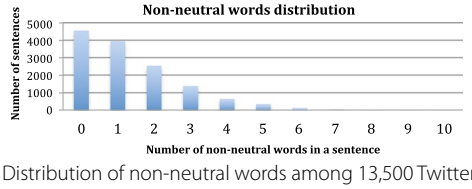

Fig. 1 Distribution of non-neutral words among 13,500 Twitter messages

Number of words	Values of *p*					
	2.00	**2.5**	**3**	**3.5**	**4**	**4 5**
1	0.59	0.62	0.66	0.69	0.72	0.74
2	0.72	0.77	0.82	0.87	0.91	0.96
3	0.82	0.89	0.96	1.02	1.09	1.15
4	0.91	1.00	1.09	1.17	1.26	1.34
5	1.00	1.11	1.21	1.32	1.43	1.54

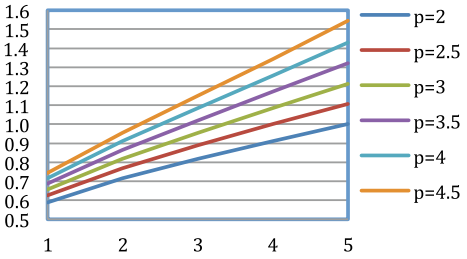

Fig. 2 Values of the coefficient presented in Eq. 5 for different numbers of non-neutral words and different values of *p*

Number of words	Average sentiment				
	20	40	60	80	100
1	14	27	41	55	69
2	17	35	52	69	87
3	20	41	61	82	100
4	23	47	70	94	100
5	26	53	79	100	100

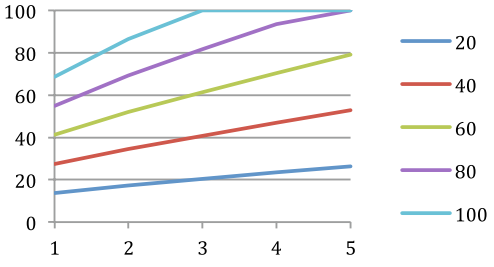

Fig. 3 Values of F_P/F_N obtained for different numbers of words and different values of the average sentiment

intensity of the positive or negative sentiment was then provided as an output.

The same formula Eq. 6 has been applied to combine pieces of evidence for all positive and negative words in case that a message contains mix sentiment. The combination formula is only applied for words with the evidence greater than 0.5. We assumed that probabilities lower than 0.5 should not be considered as evidence. While combining pieces of evidence, the max and min possible values were considered as 1 and −1 rather than 100 and −100, as given in Eq. 7.

$$e_P = min\left\{\frac{A_P}{2 - log(3.5 \times W_P)},\ 1\right\}$$
$$e_N = max\left\{\frac{A_N}{2 - log(3.5 \times W_N)},\ -1\right\} \quad (7)$$

where e_P and e_N represent overall positive and negative evidence in a sentence. Positive and negative evidence were combined separately and the outputs were considered as the final evidence that the message is positive or negative. These two values were taken under consideration in the sentiment classification process.

Sentiment classification

For a given message, in the first step of the classification process, all evidence and sentiment values are combined by using Eqs. 6 and 7. Following this, the decision process is performed as presented in Fig. 4.

The *final Sentiment* function validates the value of F_P/F_N and e_P/e_N. Depending on if the absolute value of the sentiment is greater than 25 or the absolute value of the evidence is higher than 0.5, it returns the sentiment or 0. If there are only positive words in the message, the final value of the sentiment is selected based on F_P and e_P only. The same happens if there are only negative words in the message. In case when there is a mixture of positive and negative words, the message is classify as positive or negative, depending on which, positive or negative, words are stronger. First, the difference between positive and negative evidence is calculated. If one piece of the evidence is much higher than the other (greater than 0.1) then the positive or negative sentiment is returned, respectively. In case when there is no evidence available or they do not differ strongly enough from each other, the final decision is made based on the difference between positive and negative sentiment. If the positive sentiment is greater than the negative sentiment the sentence is classify as positive and vice versa.

Empirical evaluation

The purpose of this study was to evaluate the performance of the new lexicon-based sentiment analysis algorithm within the domain of security and social media

```
IF (W_N == 0)
    RETURN finalSentiment(F_P,e_P)
ELSE IF (W_P == 0)
    RETURN finalSentiment (F_N,e_N)
ELSE {
    IF (F̄_P − F̄_N > 0.1)
        RETURN finalSentiment (F_P,e_P)
    ELSE IF (F̄_N − F̄_P > 0.1)
        RETURN finalSentiment (F_N,e_N)
    ELSE {
        IF (F_P + F_N > 0)
            RETURN finalSentiment (F_P,e_P)
        ELSE IF (F_P + F_N < 0)
            RETURN finalSentiment (F_N,e_N)
        ELSE
            RETURN 0
}}
```

```
finalSentiment (F,e) {
    if (|F| > 25  || |e| > 0.5)
    RETURN F
ELSE
    RETURN 0
}
```

Fig. 4 Sentiment classification process

analytics. In addition, we evaluated how the algorithm performed with longer messages, such as movie reviews. The same lexicon has been applied with both of the data sets. In the experiment we compared performances of five sentiment analysis algorithms, namely:

L general lexicon based technique that includes negation and intensification. Instead of application of Eq. 6 it simply sums sentiment's values of all positive and negative words within a sentence. It further classifies messages as positive, negative or neutral if the obtain value is positive, negative or equal zero, respectively.

LN performs in the same manner analysis as L, however, instead of summing it applies Eq. 6 to combine the sentiment's values of positive and negative sentiment.

LNS performs LN for each sentence within a message and calculated overall positive/negative sentiment of a sentence as an average of the values obtained for all of the sentences within a message.

LNW performs as LN but in case of mixed sentiment within a message it applies the evidence-based function presented in Eq. 7 and follows the process from Fig. 4 to classify the message as positive or negative.

LNWS performs LNW for each sentence within a message. The process from Fig. 4 is repeated for each of the sentences. The final sentiment is calculated as an average of the values obtained for all the sentences.

All the algorithms were evaluated with two data sets. The evaluation results are presented in the two following sections. In order to provide more insight, for each dataset, the best performing method was further evaluated in term of precision, recall and F-measure.

Social media

The aforementioned techniques were evaluated with the Stanford test Twitter corpus [29]. With the Stanford train dataset that was used for generating the lexicon, the sentiment was assigned automatically based on the presence of emoticons in the messages. Therefore it is not guaranteed that the labels were determined with 100 % accuracy. As opposed to the train dataset, the Stanford test dataset was manually collected and labelled hence it is more appropriate for evaluation of the classification models' performance. It contains 177 negative, 182 positive and 139 neutral manually labelled tweets. The classification accuracy of all the algorithms described in the previous section is presented in Table 1.

In an effort to gain a better insight into the obtained results, a confusion matrix was constructed for the LNW method that obtained the best results. Table 2 presents results obtained by the LNW method applied with the Stanford Twitted dataset.

Columns in the table refer to actual sentiment of the tweets from the testing set. Rows represent the sentiment predicted by the LNW method. The diagonal represents the true positive indicating the instances, which were correctly classified by the method. Based on the confusion matrix the precision, recall and F-Measure of the method were calculated and presented in Fig. 5.

Movie reviews

The proposed sentiment analysis method was designed particularly for social media data. In the future we wish to evaluate the method with data pulled from different sources such as Facebook, where messages are longer and contain multiple sentences. However, for this work

Table 1 Classification accuracy of the five algorithms applied with Stanford Twitter dataset

Classification method	Classification accuracy
L	69.1
LN	72.6
LNS	63.1
LNW	77.3
LNWS	72.7

Table 2 Confusion matrix generated based on the results obtained by the LNW method

Assigned sentiment	Labelled sentiment		
	Positive	Neutral	Negative
Positive	127	11	8
Neutral	32	110	29
Negative	15	18	149

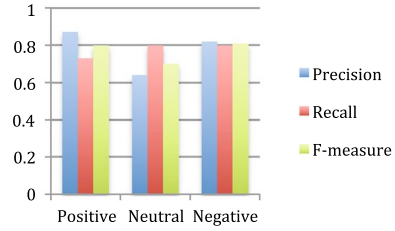

Fig. 5 Precision, recall and F-measure of the LNW method for each of the classes

Twitter corpus was the only one, manually labelled and publicly available, that we could find. In order to evaluate specific parts of the algorithm in more details, it was necessary to apply our algorithm with a set of more complex documents. Because of lack of social media data, the Internet Movie Database (IMDB) [31] with 25,000 movie reviews including 12,500 positive and 12,500 negative was selected for this purpose. The main objective for this experiment was to investigate how the proposed method performs on sentence level in comparison to document level. It was observed during the experiment with the Twitter dataset that better result could be obtained when no sentence analysis was applied. In this experiment we wanted to examine whether the same situation takes place for longer messages such as movie reviews. Given that we aimed to compare LN with LNS and LNW with LNWS, rather than to test the method in the movie domain, we used the lexicon trained with the Twitter data. The classification accuracy of all the algorithms described in the previous section is presented in Table 3.

Table 3 Classification accuracy of the five algorithms applied with Stanford IMDB

Classification method	Classification accuracy
L	67.5
LN	51.4
LNS	71
LNW	60
LNWS	74.2

Table 4 represents results obtained by the LNWS method that obtained the greatest accuracy with the IMDB dataset.

Figure 6 represents the precision, recall and F-Measure calculated for each of the classes based on the confusion matrix.

Discussion

Twitter dataset

We can observe from Table 1 that the traditional lexicon based method obtained accuracy of 69.1 % with the Twitter data set. It can be noticed that the accuracy increased when the sentiment normalisation had been applied, indicating that the normalization function expresses more accurately the intensity of the sentiment of messages in comparison to the sum function. Besides this, we can find from the results presented in Table 1 that application of the evidence-based function improves the performance of the proposed method (77.3 %). Following this, we investigated whatever better performance can be achieved while performing sentence or message level sentiment analysis. From the obtained results it can be noticed that for short messages such as tweets better accuracy was achieved for message level sentiment analysis. Finally, it can be inferred from the results that LNW was the most appropriate sentiment analysis method for Twitter data.

It can be observed, based on the results presented in Fig. 5 that in term of F-measure LNW performed the worst with the neutral messages. The precision obtained

Table 4 Confusion matrix generated based on the results obtained by the LNWS method

Assigned sentiment	Labelled sentiment		
	Positive	Neutral	Negative
Positive	8720	0	2346
Neutral	442	0	324
Negative	3338	0	9830

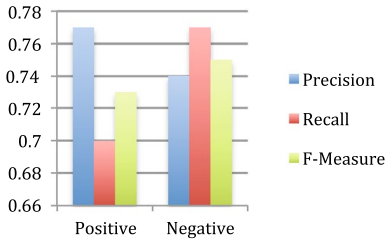

Fig. 6 Precision, recall and F-measure of the LNWS method for each of the classes

by LNW for the neutral class was only 0.64, which indicates that the most common mistake made by the method, is misclassifying positive and negative messages as being neutral. We can see from Table 2 that 32 of positive and 29 of negative messages were classified as neutral. The misclassified positive and negative tweets were assigned as neutral in majority of the cases. Only 15 positive messages were classified as negative and only 8 of negative tweets were assigned as positive.

Movies review dataset

It can be observed from the results presented in Table 3 that, as opposed to the Twitter data, application of the normalisation function with the lexicon-based method caused decrease in the accuracy from 67.5 to 51.4 %. At the same time, application of normalisations improved the performance significantly (71 %) while the sentiment analysis was performed on sentence level. This indicates that the normalisation process is more appropriate for short messages or sentences, rather than a long document. Following this, we can see that LNW and LNWS achieved 60 and 74.2 % accuracy, respectively. This shows that application of the evidence function improve the performance only in the case of sentence level sentiment analysis. Lastly, it can be inferred from the obtained results that the LNWS method was the most accurate while applied with the IMDB dataset.

It can be noticed from Table 4 that the misclassified positive reviews were more often assigned as negative (3338) rather than neutral (442). Similarly, negative reviews were more often misclassified as positive (2532) than as neutral (324). In term of precision and recall, the LNWS method performed on a similar level for both, positive and negative classes.

Following the aforementioned evaluation of the new lexicon based approach, we can conclude that for short messages, such as tweets, the method performs better on document level (LNW). For longer messages, on the other hand, the most optimal results are obtained when the method is performed on the sentence level (LNWS).

Case study: English defence league

English defence league

The English defence league (EDL) is a right wing political organisation that opposes what is considered to be the "Spread of Islamism in the United Kingdom" (http://www.englishdefenceleague.org). EDL was formed in 2009 and its principal activities have been regular street demonstrations in English and Welsh towns and cities. In this manner the group attempts to influence public opinion. EDL has number of opponents, such as Unite Against Fascism (UAF), that attend to counter their demonstrations. Even though it aims to demonstrate peacefully, conflicts with the counter demonstrators often led to street violence, anti-social behaviours and arrests. Due to the high likelihood of violence there is usually heavy policing required during EDL or opposing demonstrations. The cost of policing these demonstrations is estimated to be from £300,000 to £1 million for an event. In the past 5 years, a number of EDL and opposing demonstrations took place in England. Some of them were very peaceful without any major incidents. A few of them, however, required a large police presence.

We selected the EDL related events as a case study for our work. We aim to investigate the relation between negative sentiment of messages related to the events being posted on Twitter and the amount of disorder during the demonstrations. For this purpose we selected four EDL events described below.

20th July 2013 Birmingham

The Birmingham demonstration was organized by EDL and violent disorder with a number of clashes between EDL supporters, anti-fascist protesters and police was reported in the press. Smoke bombs, stones and bottles were thrown at the police as the EDL and the opponents gathered in the city centre for simultaneous demonstrations. According to the Birmingham Mail,[1] close to 50 people have been charged by West Midlands Police for criminal damage and assault relating to the protests.

6th February 2014 Slough

The Slough March was organized by EDL and it involved a number of counter protests. The two opposing demonstrations passed off without incident for the police. Only a very small amount of disorder broke out during the March. The local police commander for Slough said[2]: "I am pleased that these demonstrations have passed off without major incident. Disruption was kept to a minimum and we are grateful for the support received from local communities."

27th April 2014 Brighton

The Brighton demonstration was organized by the 'March for England' (MfE) organisation. During the protest the police were trying to separate 150 nationalists from more than a thousand anti-fascists demonstrators. This was considered as one of the largest police operations in Brighton. A number of violent clashes between members of each group took place followed by 27 arrests.[3]

[1] http://www.birminghammail.co.uk/news/local-news/violence-at-edl-birmingham-rally-5165256.

[2] http://www.bbc.co.uk/news/uk-england-berkshire-25999527.

[3] http://www.theargus.co.uk/news/11175736.Violent_clashes_as_March_for_England_returns_to_Brighton/?ref=var_0.

10th May 2014 Rotherham

The Rotherham demonstration was organized by EDL and it involved hundreds of people marching through the town centre. Even though a large group of UAF members was holding a counter-protest at the same time, police said the event saw minimal disruption and no disorder.[4]

Public sentiment for EDL

The focus of this study has been directed towards the analysis of the relationship between public sentiment and tension of EDL related events. In our work we attempted to investigate if the public sentiment regarding the EDL can be applied to predict (to a certain extent) the level of disruption during the event. As the first step we decided to consider Twitter as a data source, given that it is considered as the most popular social media channel. Our goal was to analyse the negative sentiment of all the messages related to EDL that had been posted on Twitter prior to each of the four events mentioned in the previous sub-section. Following this, we were able to observe if there is any relationship between the negative sentiment and the level of disruption and disorder.

For the purpose of this study we gathered data from 24 days (6 days prior to each of the events). The data was gathered through the RepKnight platform. All the obtained tweets were associated to EDL and they were identified through keyword searches. The data is summarized in Table 5 below.

For each tweet, the lexicon-based sentiment analysis algorithm introduced in "Empirical evaluation" was applied. We decided to decrease the size of the sentiment lexicon and make it more domain-specific by removing irrelevant words. For this purpose we applied a data set with 1 million EDL related tweets. For each word from the general lexicon we calculated its frequency occurred in the corpus. It appeared that only a subset of 1500 words out of 4000 had been used. From this observation, we reduced the size of the lexicon from 6000 to 1500 words. Given that we are interested in predicting levels of violence and disorder during public events, we take under consideration only the negative sentiment. As a result of the sentiment analysis process, each tweet was assigned with a normalised value from a range 0–100, where 100 represents the greatest negative sentiment's value. The two factors that we intend to analyse were the number of negative tweets posted during 6 days prior to the event and the level of negative sentiment within these messages. The data selected for each day was first analysed separately and then aggregated. Tweets selected from each day were grouped into five categories related to the strength of the negative sentiment (0–20, 20–40, 40–60, 60–80 and 80–100). Messages from the categories 0-20 and 80-100 are considered to be the least and the most negative, respectively. Each of the figures below presents two diagrams representing data related to one of the four events from Table 5. The diagrams on the left hand side (doughnuts) present all the data selected during the 6 day period. Each of the doughnuts demonstrates the distribution of the negative tweets over the five categories. Each category is represented by a different colour. The diagrams on the right hand side present the distribution of the tweets from each category over the 6 days prior to the event.

Table 5 EDL related tweets

Period	Event	Number of tweets
14–19 July	Birmingham	38,408
26–31 January	Slough	7662
21–26 April	Brighton	16,056
4–10 May	Rotherham	29,104

Following the information that has been provided in "Public sentiment for EDL" we can infer that the Birmingham EDL demonstration was the most violent one. It caused the highest level of disorder and was followed by the highest number of arrests. Based on the results presented in Fig. 7 it can be noticed that the Birmingham demonstration obtained the highest attention of Twitter's users comparing to the other events. The number of negative messages posted during the 6 days prior to the protest in Birmingham (18,038) is almost three times higher than those in Brighton (5558) and Rotherham (5352). It can be found that the number of the most negative tweets (from categories 60–80 and 80–100) is greater than the number of all negative messages gathered for the other events. The second demonstration that caused street violent and disorder was the MfE in Brighton. The other two events, namely EDL Slough and EDL Rotherham revealed minimal disruptions and can be considered as peaceful. We can observe from Fig. 8 that the protest in Slough obtained the least attention from Twitter's users. The number of negative messages and the level of the negative sentiment were much lower comparing to the two demonstrations where violence was reported. The significant difference that can be found between the graphs presented in Figs. 7 and 8 indicates that there is some correlation between the level of negative sentiment around a demonstration and the level of tension during the event. The relation between the negative sentiment and the degree of violence is not, however, noticeable from Figs. 9 and 10. Both of the events, EDL Brighton and EDL Rotherham, obtained similar amount of negative tweets, namely 5558 and 5352 respectively. In the

[4] http://southyorks.police.uk/news-syp/protest-rotherham-town-centre.

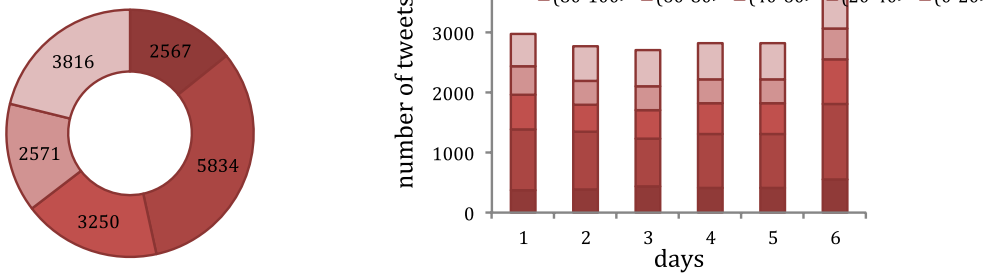

Fig. 7 Negative sentiment analysis results for EDL from the period 14–19 July 2013 prior to the EDL demonstration in Birmingham. Total number of negative tweets collected is 18,038

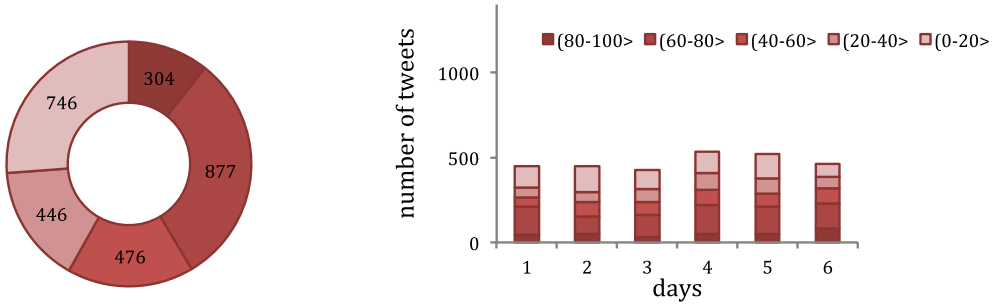

Fig. 8 Negative sentiment analysis results for EDL from the period 26–31 January 2014 prior to the EDL demonstration in Slough. Total number of negative tweets collected is 2849

same time, the event in Brighton was violent while the demonstration in Rotherham was peaceful.

The obtained results suggest that the negative sentiment around an event can be, to a certain extent, applied as an indicator for the level of disorder. Such a tool could be useful as an additional support for the police services while planning the resources around safeguarding of public events. Social media is the easiest and fastest way to source and identify what people are saying and how they are feeling about different events, for example. Analysis of negative sentiment of messages related to a public event provides information related to the state of mind of people that are going to attend or are attending the event. Such an analysis can be applied for prediction as well as monitoring disorder during public events.

Conclusions and future work

In this work we presented a new approach to lexicon-based sentiment analysis of Twitter messages. In the new approach, the sentiment is normalised, which allows us

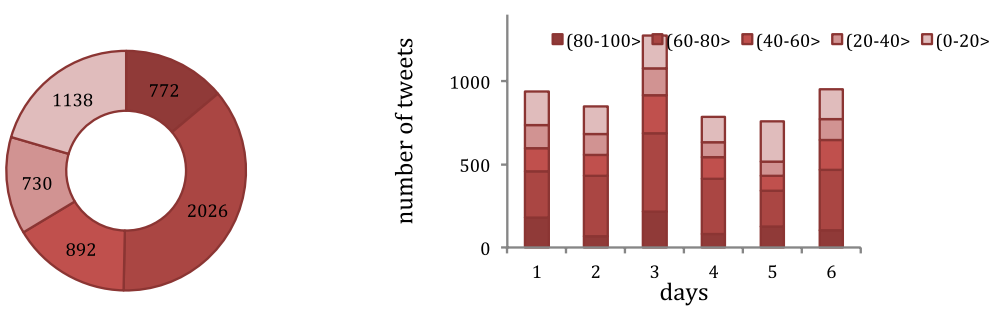

Fig. 9 Negative sentiment analysis results for EDL from the period 21–26 April 2014 prior to the MfE demonstration in Brighton. Total number of negative tweets collected is 5558

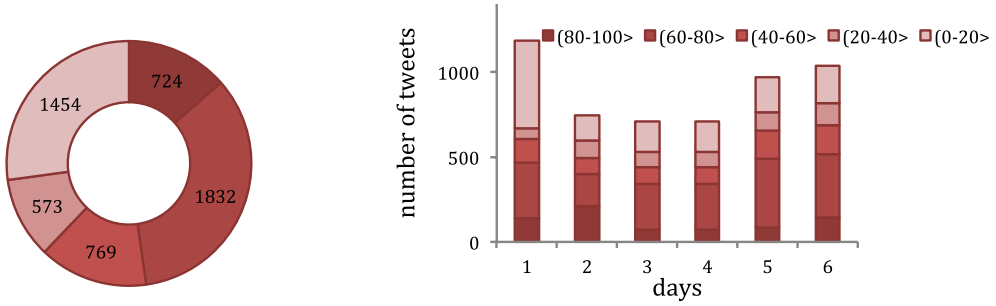

Fig. 10 Negative sentiment analysis results for EDL from the period 4–10 May 2014 prior to the EDL demonstration in Rotherham. Total number of negative tweets collected is 5352

to obtain the intensity of sentiment rather than positive/negative decision. A new evidence-based combining function was developed in an effort to improve performance of the algorithm in the cases where a mixed sentiment occurs in a message. The evaluation was performed with the Stanford Twitter test set and IMDB data set. It was found from the results that the two new functions improve performance of the standard lexicon-based sentiment analysis algorithm. It could be noticed that the method is more appropriate for short messages such as tweets. When applied with long documents the method performed significantly better on the sentence than on the document level. Following this, our intention was to investigate the relationship between the amount and the level of negative sentiment related to a public demonstration and the level of violence and disorder during the event. In other words, we aimed to ascertain if sentiment analysis could be applied as a supportive tool while predicting a level of disruption prior to public events. As a first step in this study we decided to examine Twitter as a source of data. Four different demonstrations were selected and the negative sentiment related to these events was analysed over 6 days prior to each event.

Following the case study and a number of analyses we were able to reveal that there was a relationship to some extent between the negative sentiment and the level of disorder during the EDL events. Further research is however required in this area in an effort to provide more accurate findings and conclusions. At the current stage we can, however, conjecture that sentiment analysis of social media content can provide valuable, security-related information regarding some upcoming public events. In the next step we wish to collect more data related to public events and further investigate the relationship between negative sentiment and the level of violence and disorder during events. Following this, we aim to develop a predictive model that can be used by police services as a single tool to help indicate violence propensity.

In future work we wish to focus more on multilingual sentiment analysis. Given that data pulled from social media are created by users from all over the globe, there is a consequent demand to perform sentiment analysis in more than just one language. The most challenging problem while trying to translate sentiment lexicon in a different language is inflection and conjugation of words applied in some of the languages. Unlike in English, some languages make use of grammatical gender and plural. Following this, verbs, nouns and adjectives are inflected for person or number and verbs are marked for tense. For example, while in English the verb "love" can be used in 4 different forms (love, loved, loving, loves), in Polish language there are 20 different forms depending of tense and person. Besides this, the adjective "nice", for example, in Polish language can be used in 5 different forms. Consequently, it would be very inefficient to include all the different forms of words in the lexicon, especially when talking about real time analysis. In some preliminary work we were able to demonstrate that by application of an appropriate string similarity function it is possible to perform sentiment analysis with the lexicon containing only regular form of words. Another important issue while translating a lexicon into another language is disambiguation. It is important to ensure that for ambiguous words, the appropriate meanings have been translated and included into new lexicon. Consequently, an automatic translation may not provide the desired results. In our work, semi-automatic translation has been applied where all ambiguous words were translated manually. We were able to show that by translating words from the English lexicon into regular Polish and Portuguese words and by application of a string similarity function, the sentiment analysis of Polish/Portuguese tweets can be performed on a similar level of accuracy as for the English language. At the same time, some preliminary experiments demonstrated that the proposed method could be easily adapted to languages such as Malay, where no inflection or conjugation is being applied to the words.

In the future work we intend to evaluate the multilingual version of the method in more details.

Authors' contributions
AJ was the lead researcher for this work and undertook the design and development of the sentiment analysis algorithm, data analysis and preparation of the manuscript. MM and YB supported the design, from a theoretical perspective, of the lexicon based approach and supported the preparation of the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This work undertaken between Ulster University and Repknight Ltd gratefully acknowledges support from Innovate UK's Knowledge Transfer Partnership, project No. KTP009125.

Competing interests
The authors declare that they have no competing interests.

Received: 13 March 2015 Accepted: 19 November 2015
Published online: 09 December 2015

References

1. LexisNexis® Risk Solutions (2012) Survey of law enforcement personnel and their use of social media in investigations. http://www.lexisnexis.com/investigations
2. Grabner D, Zanker M, Fliedl G, Fuchs M (2012) Classification of customer reviews based on sentiment analysis. In: proceeding of International Conference on Information and Communication Technologies in Tourism, pp 460–470
3. Krauss J, Nann S, Simon D, Fischbach K, Gloor P (2008) Predicting movie success and academy awards through sentiment and social network analysis. In: Proceedings of European Conference on Information Systems (ECIS)
4. Asur S, Huberman BA (2010) Predicting the future with social media. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp 492–499
5. Hu Y, Wang F, Kambhampati S (2013) Listening to the crowd: automated analysis of events via aggregated twitter sentiment. In: Proceeding of International Joint Conference on Artificial Intelligence, pp 2540–2646
6. Garcia A, Gaines S, Linaza MT (2012) A lexicon based sentiment analysis retrieval system for tourism domain. Expert Syst Appl Int J 39(10):9166–9180
7. Xu J, Zhu X, Bellmore A (2012) Fast learning for sentiment analysis on bullying. In: Proceeding of International Workshop on Issues of Sentiment Discovery and Opinion Mining
8. Mittal A, Goel A (2013) Stock prediction using twitter sentiment analysis. In: Proceeding of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology
9. Cataldi M, Caro LD, Schifanella C (2010) Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceeding of International Workshop on Multimedia Data Mining
10. Lampos V, Bie TD, Cristianini N (2010) Flu detector—tracking epidemics on Twitter. In: Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases, pp 599–602
11. Sadilek A, Kautz H, Silenzio V (2012) Predicting disease transmission from geo-tagged micro-blog data. In: Proceedings of AAAI Conference on Artificial Intelligence
12. Glass K, Colbaugh R (2011) Web analytics for security informatics. In: Proceedings of European Intelligence and Security Informatics Conference, pp 214–219
13. Tyshchuk Y, Wallace W, Li H, Ji H, Kase S (2014) The nature of communications and emerging communities on Twitter following the 2013 Syria Sarin Gas Attack. In: Proceeding of IEEE JISIC
14. Kong Q, Mao W, Dajun Zeng D, Wang L (2014) Predicting popularity of forum threads for public events security. In: Proceeding of IEEE JISIC, pp 99–106
15. Spitters M, Eendebak PT, Worm DTH, Bouma H (2014) Threat detection in Tweets with Trigger patterns and contextual cues. In: Proceeding of IEEE JISIC, pp 216–219
16. Cano AE, He Y, Liu K, Zhao J (2013) A weakly supervised bayesian model for violence detection in social media. In: Proceeding of IJCNLP
17. Sakaki T, Toriumi F, Matsuo Y (2011) Tweet trend analysis in an emergency situation. Proc ACM SWID 3:1–3:8
18. Doan S, Vo BKH, Collier N (2011) An analysis of Twitter messages in the 2011 Tohoku Earthquake. eHealth 58–66
19. Colbaugh R, Glass K (2011) Agile sentiment analysis of social media content for security informatics applications. In: Proceedings of European in Intelligence and Security Informatics Conference, pp 327–331
20. Colbaugh R, Glass K (2013) Analysing social media content for security informatics. In: Proceeding of European Intelligence and Security Informatics Conference, pp 45–51
21. Li W, Chen H (2014) Identifying top sellers in underground economy using deep learning-based sentiment analysis. In: Proceeding IEEE JISIC, pp 64–67
22. Westling A, Brynielsson J, Gustavi T (2014) Mining the web for sympathy: the pussy riot case. In: Proceeding IEEE JISIC, pp 123–128
23. Bermingham A, Conway M, McInerney L, O'Hare N, Smeaton AF (2009) Combining social network analysis and sentiment analysis to explore the potential for online radicalization. ASONAM
24. Cohen K, Johansson F, Kaati L, Mork JC (2014) Detecting linguistic markers for radical violence in social media. Terror Polit Violence 26(1):246–256
25. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Comput Linguist J 267–307
26. Tong RM (2001) An operational system for detecting and tracking opinions in on-line discussions. In: Working Notes of the SIGIR Workshop on Operational Text Classification, pp 1–6
27. Turney P, Littman M (2003) Measuring praise and criticism: inference of semantic orientation from association. ACM Transact Inform Syst J 21(4):315–346
28. Esuli A, Sebastiani E (2006) SentiWordNet: a publicly available lexical resource for opinion mining. In: Proceedings of language resources and evaluation (LREC)
29. Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. Technical Project, Stanford Digital Library Technologies Project
30. Sauri R (2008) A factuality profiler for eventualities in text. PhD Thesis
31. Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. In: Proceedings of ACL, pp 142–150
32. Jurek A, Bi Y, Mulvenna MD (2014) Twitter sentiment analysis for security-related information gathering. In: Proceedings of IEEE JISIC, pp 48–55