Electron Commer Res (2018) 18:181–199
https://doi.org/10.1007/s10660-017-9257-8

# A model for sentiment and emotion analysis of unstructured social media text

**Jitendra Kumar Rout[1] · Kim-Kwang Raymond Choo[2] · Amiya Kumar Dash[1] · Sambit Bakshi[1] · Sanjay Kumar Jena[1] · Karen L. Williams[2]**

Published online: 4 April 2017


**Abstract** Sentiment analysis has applications in diverse contexts such as in the gathering and analysis of opinions of individuals about various products, issues, social, and political events. Understanding public opinion can help improve decision making. Opinion mining is a way of retrieving information via search engines, blogs, microblogs and social networks. Individual opinions are unique to each person, and Twitter tweets are an invaluable source of this type of data. However, the huge volume and unstructured nature of text/opinion data pose a challenge to analyzing the data efficiently. Accordingly, proficient algorithms/computational strategies are required for mining and condensing tweets as well as finding sentiment bearing words. Most existing computational methods/models/algorithms in the literature for identifying sentiments from such unstructured data rely on machine learning techniques with the bag-of-word approach as their basis. In this work, we

✉ Sambit Bakshi
sambitbaksi@gmail.com; bakshisambit@nitrkl.ac.in

Jitendra Kumar Rout
jitu2rout@gmail.com

Kim-Kwang Raymond Choo
raymond.choo@fulbrightmail.org

Amiya Kumar Dash
dash.amiyakumar@gmail.com

Sanjay Kumar Jena
skjena@nitrkl.ac.in

Karen L. Williams
karen.williams@utsa.edu

1 Department of Computer Science, National Institute of Technology, Rourkela, Odisha 769 008, India

2 Department of Information Systems and Cyber Security, University of Texas at San Antonio, San Antonio, TX 78249-0631, USA

use both unsupervised and supervised approaches on various datasets. Unsupervised approach is being used for the automatic identification of sentiment for tweets acquired from Twitter public domain. Different machine learning algorithms such as Multinomial Naive Bayes (MNB), Maximum Entropy and Support Vector Machines are applied for sentiment identification of tweets as well as to examine the effectiveness of various feature combinations. In our experiment on tweets, we achieve an accuracy of 80.68% using the proposed unsupervised approach, in comparison to the lexicon based approach (the latter gives an accuracy of 75.20%). In our experiments, the supervised approach where we combine unigram, bigram and Part-of-Speech as feature is efficient in finding emotion and sentiment of unstructured data. For short message services, using the unigram feature with MNB classifier allows us to achieve an accuracy of 67%.

**Keywords** Sentiment analysis · Bag-of-words · Lexicon · Laplace smoothing · Parts-of-Speech (POS) · Machine learning

## 1 Introduction

Decision making is an important human aspect, and one factor that influences one's decision making is "What other people think (or are thinking)". For example, prior to the existence of the World Wide Web, friends were asked to suggest an automobile mechanic or to explain who they were planning to vote for in elections. However, the Internet have made it possible to search and source for opinions from a vast pool of individuals outside one's personal network (e.g. from sources not directly connected to the particular individual). Furthermore, more individuals are making their opinions accessible to outsiders by means of the Internet (e.g. publishing comments and blogs on social networking sites).

Sentiment analysis, also referred to as *opinion mining*, is an approach for analyzing people's opinion, sentiment, attitude, evaluation and emotions towards an entity. An opinion is an individual's belief and is the source of an individuals assessment, judgment and evaluation of any event or idea. Liu et al. [16] conclude that others' opinions have a great impact on and provide guidance for individuals, organizations and social communities during the decision making process. Precise and timely information is generally the basis for effective decision making. Individuals often tend to refer to specialists, friends, and relatives in order to validate their opinion, which can be either positive or negative.

Sentiment analysis can be challenging to perform due to the semantical and syntactical viability of the language, and the involvement of implicit extraction of objects and their assessment by keeping intact the emotions and sentiments involved. Thus, automatic identification of sentiment needs fine grained linguistic analysis techniques as well as significant efforts in extracting suitable features for machine learning or rule-based approaches.

According to Liu [14], sentiment analysis is defined as "Given a set of evaluative text documents *D* that contain opinions (or sentiments) about an object, sentiment analysis aims to extract attributes and components of the object that have been commented on in each document $d \in D$ and to figure out whether the comments are positive, negative or neutral". Usually, an opinion is articulated by an individual (opinion holder) who conveys a perception (positive, negative, or neutral) about an entity (target object e.g. person, item, organization, event, service, etc.).

In our increasingly connected society, a large proportion of our population is connected via the social networks, either directly or indirectly. Therefore, anyone can freely express his/her views/opinions without unveiling his/her true identity and without the fear of undesirable consequences (e.g. persecution). It allows people with hidden agenda and malicious intention to masquerade their opinions [10] as independent members of society, leading to the posting of fake reviews, the discrediting of legitimate products and giving of a false impression which leads to opinion spamming or astroturfing [27–29]. Another challenge to sentiment analysis is that users typically specify mixed feelings. Thus, it is not surprising that sentiment analysis remains a topic of ongoing interest.

While there has been a fair amount of research on how sentiments are conveyed in classification of online reviews [2, 3, 7, 13, 15, 18, 26, 31, 32, 35, 37, 39, 40] and news articles [8, 11, 12, 17, 19, 21, 30], sentiment analysis of informal language and microblogs is under-studied. Unlike product reviews and lengthy comments, the shorter length of microblogging text and the use of informal language (e.g. abbreviation and emoticons) complicate the analysis. Resources, such as sentiment lexicons, can be useful for sentiment analysis of online reviews and news articles. However, *whether these resources and features are helpful for sentiment analysis of microblogs (e.g. tweets) and informal language* remains to be investigated. This is the focus of this paper.

In this paper, we extend the previous approach in [33], which comprises both unsupervised and supervised techniques to tweets obtained from the twitter public domain and movie reviews. Specifically, in the unsupervised approach, we use the SentiWordNet lexicon; performance of this approach is assessed by the number of tweets that are misclassified. In the supervised approach, we use Naive Bayes classifier for emotion analysis of tweets and short message service (SMS). The performance of the classifier is then evaluated in terms of precision, recall F-measure, and accuracy.

The remainder of the paper is organized as follows: Sect. 2 reviews existing literature on sentiment analysis. Sections 3 and 4 describe the data collection and per-processing methods, and the unsupervised and supervised approaches for sentiment analysis, respectively. Section 5 concludes the findings of this paper.

## 2 Related work

In recent years, much research has been conducted on sentiment analysis to develop systems that are more reliable and provide better accuracy. The core task of sentiment analysis is the automatic identification of opinionated text in documents [20]. Previous research used both rule based and statistical machine learning approaches for opinion mining and sentiment analysis [24]. In this section, we briefly discuss some techniques on sentiment analysis and their applications.

Ibrahim et al. [34] presented a detailed survey of different techniques used for opinion mining and sentiment analysis. Turney [38] suggested an unsupervised algorithm, which uses semantic orientation of the phrases for classification of reviews. The lexicon-based approach determines the polarity or sentiment using some function of opinion words in the document or sentence [4, 36]. Esuli et al. [5] developed the SentiWordNet lexicon which contains opinion strength for each term. The feasibility of SentiWordNet lexicon for sentiment classification of documents has been assessed by Ohana [22]. Hamouda et al. [9] used the SentiWordNet Lexicon to classify reviews. A dictionary based technique is proposed by Fei et al. [6] to identify aspects of a review by considering adjectives only.

Pang and Lee [23, 25] used NB, ME and SVM for sentiment analysis of movie reviews by taking into account some special features like unigrams, bigrams, and a combination of both (i.e. unigram and bigram), including POS and positional information with unigram, and adjectives. It was evident from their experiment that feature presence provides greater accuracy than feature frequency. For small feature space, NB performs better than SVM. However, when feature space is increased, SVM outperforms NB classifier. Bikel et al. [1] implemented a subsequence kernel based voted perceptron and evaluated its performance with standard SVM. The authors observed that the increase in the number of false positives with the increase in the number of true positives is far less in aforesaid scheme when compared with the bag-of-words based SVM, where the trend is almost linear. The model shows resiliency over the fact that a smooth continuum is observed for intermediate star rating reviews even when trained only on the extreme one- and five-star rating reviews. Further, microblog sentiment analysis can be evaluated through various classifiers in two phases under two different settings. The first phase involves separation of subjective and objective documents using various classifiers. In the second phase, these filtered documents are tagged as positive or negative by the classifiers.

In this paper, we perform sentence-level sentiment identification, where tweets are classified using both unsupervised and supervised algorithms. To increase the accuracy of the classifier, the tweets are pre-processed for removal of the non-polar words, and we consider only the polar words that give sentiment. The proposed unsupervised method yields better performance for identifying sentiment of a given tweet as compared to other methods, as well as effectively classifying the tweets. We also demonstrate that the supervised approach with unigram, bigram and POS as features are effective in identifying emotion and sentiment of unstructured data. In addition, we found MNB classifier with unigram as a feature to be more effective in classifying short messages as compared to other alternative features.

# 3 Data collection and preprocessing

## 3.1 Data collection

In this research, we use tweets (up to 140 characters per tweet) as the source of our data which can be collected using *Twitter API*.

Twitter offers two APIs, namely: REST and Streaming. Streaming API differs from REST API in the sense that its connection is long-lived and offers data in near real time. In contrast, the REST API supports short-lived connections and is rate limited (i.e. one can download a restricted amount of data per day). The Twitter data, such as status updates and user information, are accessed by the REST API. However, availability of data older than a week is not supported by Twitter. In other words, REST API allows us to access accumulated data while Streaming API provides access to data as it is being tweeted.

In this work, we aim to identify the sentiment (positive,negative and neutral) of tweets for a particular product or movie. Therefore, only tweets about that that particular product or movie should be collected. However, this is not a trivial task. There seems to be no easy way of obtaining all tweets for a particular object. To retrieve tweets, Twython, a third-party API, is used. Twython is a pure python wrapper for the Twitter API, which supports both normal and streaming Twitter APIs. A total of 60,195 tweets were collected by running a script using Twython API. The corpus contains tweets about Apple, Google, Microsoft and Twitter—refer to Table 1 which provides an overview of the topicwise categorization of the corpus.

## 3.2 Training data

There are two datasets used for the training of a classifier, namely: subjective data and neutral data. Subjective data reflects positive and/or negative sentiment, while neutral data do not have any sentiment.

### *3.2.1 Subjective data*

Subjective data in this context are data that contain negative and/or positive sentiment or emoticons. While it is possible to collect sufficient negative and

**Table 1** Data by topics

| Topic | # Positive | # Neutral | # Negative | # Irrelevant | Twitter search term |
|---|---|---|---|---|---|
| Google | 2480 | 4815 | 1616 | 5280 | # Google |
| Apple | 2218 | 6102 | 4070 | 1940 | # Apple |
| Twitter | 2537 | 5245 | 2112 | 6410 | # Twitter |
| Microsoft | 2250 | 4487 | 3203 | 5430 | # Microsoft |
| Total | 9485 | 20,649 | 11,001 | 19,060 | 60,195 |

positive data in one or two consecutive days by running a script using some query entities, for convenience manually annotated positive and negative tweets are being used as the subjectivity dataset to train the classifier. As observed from Table 1, out of 60,195 tweets, a total of 11,001 tweets are negative and 9485 are positive. The fact that there are more negative tweets than positive tweets show that more people use negative sentiment than positive sentiment. There are substantially more tweets that contain both negative and positive sentiments. Tweets that contain both negative and positive sentiments are confusing because they contain both sentiments, so these tweets were annotated as neutral.

#### *3.2.2 Removing non-English tweets*

The tweets being retrieved may contain both English and non-English tweets. As the objective in this paper is to find the sentiment of tweets that are in the English language, non-English tweets were eliminated from the dataset and the classifier was trained using English tweets only. This was possible by using Google's language detection web service, which requires a reference website against which strings are compared to determine their language. The strings in this case are the tweets. The web service enables one to specify a confidence level that ranges from 0 to 1. Since Twitter data contains a lot of slang and misspellings, the confidence level was set at 0 in order to avoid the need to eliminate these (English) tweets.

#### *3.2.3 Neutral tweets*

Neutral tweets in this context are tweets that contain only factual words about a product or contain both positive and negative sentiments.

### 3.3 Data preprocessing

Preprocessing of data is essential to remove incomplete, noisy and inconsistent data. Preprocessing must be done in order to apply any data mining functionality. We employed the following preprocessing activities before applying any of the sentiment analysis approaches (i.e. lexicon based or machine learning).

- *Removing URLs*
  In general, for analysis of sentiment of the tweets, URLs cannot be considered. Let us use the sentence, "She has logged into www.ecstasy.com because she is bored," as an example. Although this sentence is negative, the presence of the word 'ecstasy' (which has a positive sentiment) in the URL may result in a neutral outcome. To overcome this limitation, the URLs are removed.
- *Filtering*
  Usually, people use words with repeated letters like 'coooool' or 'happppyyyy' to reveal intensity of expression. Since these words do not generally exist in the English language, we introduce a rule to eliminate the extra letters (i.e. a letter cannot be repeated more than three times).

- *Removing question words and stop words*
  Question words e.g. *what, which, how* etc. do not contribute to polarity. Hence, such words can be removed to ensure complexity reduction. Stop words such as *for, above, about* are also discarded as they rarely contribute to the detection of polarity.
- *Removing special characters*
  In order to resolve discrepancies during assignment of polarity, special characters like [ ], { }, ( ) should be avoided. Unless the special characters are removed, they may concatenate and make those words unavailable in the dictionary. In order to overcome this, special characters were removed.
- *Removal of retweets*
  Re-tweeting can be defined as the process of copying the tweet of another user and posting to another account. Usually, this takes place when a user likes another user's tweets. Retweets are commonly abbreviated with RT. For example, the following tweet may be considered: RT @Anum3288: Finally made a full paper box \U0001f49c # ArtLovers # paperbox # creativity # giftbox.
- *Removing hash symbols*
  Hashtags are labels used on posts in social networking sites or microblogging sites. These labels facilitate easier search for particular messages and posts and act as keywords. For example, a search on #LOST (or #Lost or #lost, because it is not case-sensitive) will produce a list of tweets related to #lost. Generally this symbol is used to indicate nouns. The hash symbols were stripped (e.g., *#tomorrow* → *tomorrow*) as they are not needed for polarity detection.

## 4 Proposed methodology

The following section describes the unsupervised and supervised methodologies used in this article for analyzing sentiment and emotion from unstructured social media text.

### 4.1 Unsupervised approach

Past research uses different lexicons for sentiment analysis of tweets. A problem with these lexicons occurs because there are certain domain specific words that are not present in these lexicons. To determine the sentiment of these words is more challenging. The proposed model does not require any data to train the classifier or any corpus to generate domain specific lexicons. In this work, we considered "excellent" to be an extremely positive word and "poor" to be an extremely negative word. The synonyms of both words were compiled to form a positive list and a negative list. NLTK POS tagger is used to find the adjectives, verbs and nouns for a given tweet. For each word in the (adjective, verb, noun) list, we used Google to find the number of hits with each word in the extremely positive list and number of hits with each word in the extremely negative list. A score was generated for each word($t$) using the following rule:

$$score(t) = \log\left(\frac{hits(t \wedge excellent) * hits(poor)}{hits(t \wedge poor) * hits(excellent)}\right) \tag{1}$$

Based on the maximum score of each word, a weight is given to each sentence/tweet. If the weight of the sentence is positive, then the sentence is classified as positive. Otherwise, it is classified as negative (Fig. 1).

### 4.2 Supervised approach

A supervised learning algorithm requires training data as input and generates an inferred function that can be used for mapping new samples. Ideally, this function can map accurately the unseen examples to class labels as they are. This is approximated by generalization of the training examples so that the learning algorithm can label unseen situations in a reasonable way. This is the phase which requires the most amount of time and effort.

A three-step process was used to conduct experiments across various machine learning algorithms in order to identify underlying factors affecting the performance.

- *Step1* Preprocessing the training data;
- *Step2* Feature extraction and data representation; and
- *Step3* Training the classifier and testing with different machine learning algorithms.

#### 4.2.1 Feature extraction

Representation of data is of utmost importance prior to any machine learning application. Sometimes, information contained in the input data to a learning

```
great battery life perfect size but a tid bit quieter than i would like
['great', 'battery', 'life', 'perfect', 'size', 'tid', 'bit', 'quieter', 'would'
, 'like', '']
[('great', 'JJ'), ('battery', 'NN'), ('life', 'NN'), ('perfect', 'NN'), ('size',
 'NN'), ('tid', 'VBD'), ('bit', 'NN'), ('quieter', 'NN'), ('would', 'MD'), ('lik
e', 'VB'), ('', '-NONE-')]
List of adjective:
['great']
list of noun:
['battery', 'life', 'perfect', 'size', 'bit', 'quieter']
list of verb:
['tid', 'like']
626914000000000
664350000000000
370720000000
694220000000
495838000000000
671560000000000
-2.32583732244
```

**Fig. 1** Snapshot of the tweet weight of proposed model algorithm may either be too vast for processing in feasible time or too redundant to be relevant. At this point, a reduction may be performed to reduce the input data to a reduced set of features. This process is called *feature extraction*. The features extracted from input data are expected to contain the relevant information, so that the desired assignment can be performed by utilizing this reduced representation rather than the complete initial information.

In order to build a model using machine learning algorithms, both the training and test data must be represented in some way. Most machine learning algorithms used BoW representation as their feature. In machine learning, feature refers to some attributes that are thought to capture the pattern of the data, and the entire dataset must be depicted in terms of these features before it is supplied to any machine learning algorithm. A variety of features like unigram, bigram, POS, syntactic and semantic feature are used for sentiment analysis.

In BoW representation, a tweet is represented as the bag of its words and phrases, disregarding grammar and even word order but keeping multiplicity. But in case of set-of-tweets representation of a Twitter post if a word occurs twice, it will only be presented once regardless of how many times it is found in the Twitter post. In this section, bag-of-word and feature representation are used for different purposes at different stages.

- *Attribute selection:* This is the method of extracting features of the data to be delineated. In machine learning algorithm for representing data instances, attribute selection is generally the first task to be performed. Once attributes are selected, the training or testing data will be described using these attributes. Thus, attributes are the features. For attribute selection, the entire dataset is being converted into unigrams/bigrams/trigrams or a combination of any of them. Bag-of-unigrams is equivalent to BoW. During attribute selection, we need to think about which words to exclude from being selected as attributes. As there are certain words which do not contribute to polarity detection, such attributes were removed. This is done by using stop words.
- *Instance representation:* Once attributes are chosen, the data must be represented in terms of these attributes. Now these attributes are called features. A decision about whether to utilize unigram presence or unigram frequency, bigram presence or bigram frequency, trigram presence or trigram frequency, or a mixture of unigram and bigram presence or frequency, and so on must be made.

### *4.2.2 Training and testing the classifier*

For training and testing the classifier, we need to determine the feature vector. The formation of a good feature vector is highly significant in implementing a classifier. A feature vector is a vector that contains information describing an object's important characteristics. The feature vectors used for the proposed approach contain features such as unigram presence, bigram presence or combination of both. The entire feature vector will be a combination of feature words of each data instance (Twitter post). The classifier was trained using these feature vectors. For training the classifier, three different machine learning algorithms were used which are described in the following section.

For testing a tweet, it is necessary to figure out the feature words which result in another pattern of feature vector. This feature vector acts as an input to the model for learning, and based on the learning, the classifiers predict/anticipate the sentiment of the tweets. In this experiment, MNB classifier with Laplace smoothing is used for classifying emotion of tweets and sentiment of SMS.

#### *4.2.3 Experimental setup*

People tend to use *hashtags* to express their sentiment or emotions, so these hash-tagged words are a good indication of sentiment and emotion. We used these hashtags to add more data to our machine learning algorithm and form a large emotion dataset. There are seven categories of emotions, namely: *anger, love, fear, joy, sadness, surprise, thankfulness*. In this work, these seven emotions were used as keywords to collect the tweets. A snapshot of the tweets retrieved using *hashtag* is shown in Fig. 2. Figures 3, 4, and 5 represent the volume of tweets retrieved with respect to time for *#anger*, *#sad* and *#surprise*, respectively.

#### *4.2.4 Evaluation metrics*

To evaluate the performance of individual classifiers, the following evaluation metrics were used:

- *Accuracy*

$$\begin{aligned} Accuracy &= \frac{No.\,of\ \ correctly\ \ labeled\ \ tweets}{Total\,No.\,of\ \ tweets\ \ in\ \ the\ \ test\ \ set} \\ &= \frac{TP + TN}{P + N} \end{aligned} \tag{2}$$

| Service | Term | Username | Name | Update | Link | Location | Followers | Friends | Time(PDT) |
|---|---|---|---|---|---|---|---|---|---|
| twitter | #joy | mariellabella | mariella | #pasta #meatballs #food #lunch #happy #mom #joy #yummy #handmac | http://twitter. | Trinidad | 105 | 445 | 07-04-2015 09 |
| twitter | #joy | KariJoys | Kari Joys MS | RT @DiRiseborough: #Letgo of the weight & find your #JOY on the | http://twitter. | Spokane, WA | 23425 | 24309 | 07-04-2015 09 |
| twitter | #joy | AllGodsThings | Omnia Dei | Alleluia! "@RadiateLA: #Easter is all about #joy. Radiate it! #RadiateLA | http://twitter. | Caribbean | 266 | 338 | 07-04-2015 09 |
| twitter | #joy | Rastamon86 | fiestas EDM madrid | #Zoologicoclub #capital #joy #marcoAldani #madriz #salír #wekeend #d | http://twitter. | Madrid | 1714 | 1366 | 07-04-2015 09 |
| twitter | #joy | debsie301261 | Deborah Myerscough | RT @ramblingsloa: " #Joy attracts #Joy" - Rhonda Byrne @thesecret @S | http://twitter. | | 525 | 581 | 07-04-2015 09 |
| twitter | #joy | upfortomorrow | Tamara | RT @headquarters: I'm gonna root root root for BOTH teams! #Opening | http://twitter. | | 18 | 82 | 07-04-2015 09 |
| twitter | #joy | wtjohnson01 | Whitney Johnson | The little spaces we create for peace. #peaceful #spring #littlethings #l | http://twitter. | | 292 | 620 | 07-04-2015 09 |
| twitter | #joy | ashrafxd | ashraf | RT @pixodentist: #FriFotos Blossom the #joy of Spring. http://t.co/7pA | http://twitter. | God`s own country / ma | 314 | 1046 | 07-04-2015 09 |
| twitter | #joy | LessysWorld | Lessy | And you should too! □□ #Joy #TrueJoyComesThroughChrist □ https://t. | http://twitter. | Brooklyn, NY | 717 | 650 | 07-04-2015 09 |
| twitter | #joy | eraivy | eraivy | RT @JohnSilkauskas: Start #singing start humming, start #worshiping ri | http://twitter. | kualalumpur | 936 | 810 | 07-04-2015 09 |
| twitter | #joy | NanaJenLaw | Nátalie Tavares | RT @Newsin01887: @JenLawUS @olv #Joy is filming RIGHT NOW at the | http://twitter. | RJ, Brasil | 471 | 566 | 07-04-2015 09 |
| twitter | #joy | patlejarde | Patricia Lejarde | luv u girlz @erikalaylo @ivymancilla @AsiRaihana @_coleenqb @jshxa | http://twitter. | PH | 743 | 285 | 07-04-2015 09 |
| twitter | #joy | TheJoyTrain | #JoyTrain | RT @DiRiseborough: #Letgo of the weight & find your #JOY on the | http://twitter. | Global | 2269 | 2414 | 07-04-2015 09 |
| twitter | #joy | lwrenci | amendoa | RT @Newsin01887: @JenLawUS @olv #Joy is filming RIGHT NOW at the | http://twitter. | | 1944 | 776 | 07-04-2015 09 |
| twitter | #joy | bydefault904 | Drizzy | @adamcarolla @Jokoy you don't understand Jo. We have a serious bro | http://twitter. | NE Florida | 109 | 318 | 07-04-2015 09 |
| twitter | #joy | kkax_ | ลพดอmine | RT @parkkim9699: [CAP] 150404 #JOY □http://t.co/pyalllp09N ©Studio_ | http://twitter. | s k r 2 1 。 | 6039 | 131 | 07-04-2015 09 |
| twitter | #joy | mathewdennis | Dennis Mathew | RT@Samsamum Focus for today Making peace with my complexity - #l | http://twitter. | Canada eh! | 26284 | 26089 | 07-04-2015 09 |
| twitter | #joy | TheJoyTrain | #JoyTrain | RT @KariJoys: @adservio Transform your #Anxiety to #Joy!#Spokane #J | http://twitter. | Global | 2269 | 2414 | 07-04-2015 09 |
| twitter | #joy | TheJoyTrain | #JoyTrain | RT @KariJoys: Real, Inspiring #Stories~People You May Know~Found #J | http://twitter. | Global | 2269 | 2414 | 07-04-2015 09 |
| twitter | #joy | xolovebeam | บีมสล็อต | RT @babyvelvet_th: [HQ] 150328 #JOY at Fansign Event In Busan ©man | http://twitter. | | 625 | 1546 | 07-04-2015 09 |
| twitter | #joy | TheJoyTrain | #JoyTrain | RT @KariJoys: Welcome to the #JoyTrain community! Just RT other twe | http://twitter. | Global | 2269 | 2414 | 07-04-2015 09 |
| twitter | #joy | TheJoyTrain | #JoyTrain | RT @KariJoys: #NewBook> "Who Stole Your #Joy?" Inspiring #hope! | http://twitter. | Global | 2269 | 2414 | 07-04-2015 09 |
| twitter | #joy | xolovebeam | บีมสล็อต | RT @babyvelvet_th: [OFFICIAL] 150407 #JOY & #JIMIN at After Sch | http://twitter. | | 625 | 1546 | 07-04-2015 09 |

**Fig. 2** Snapshot of the tweets collected via third-party API

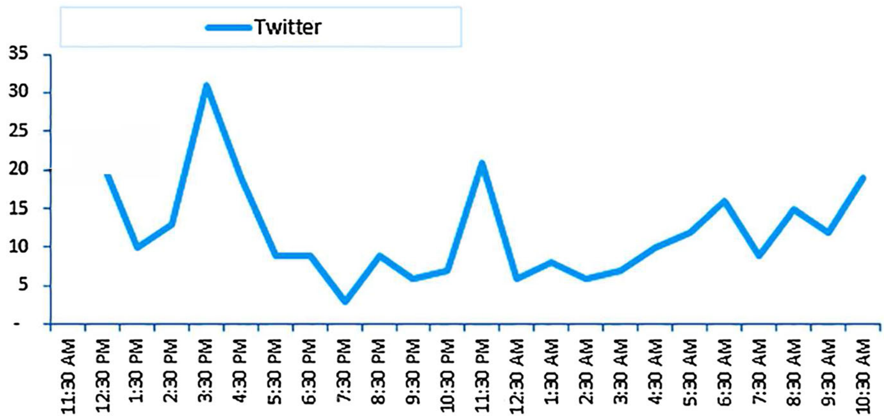


**Fig. 3** Snapshot of volume of tweets with respect to time for # anger

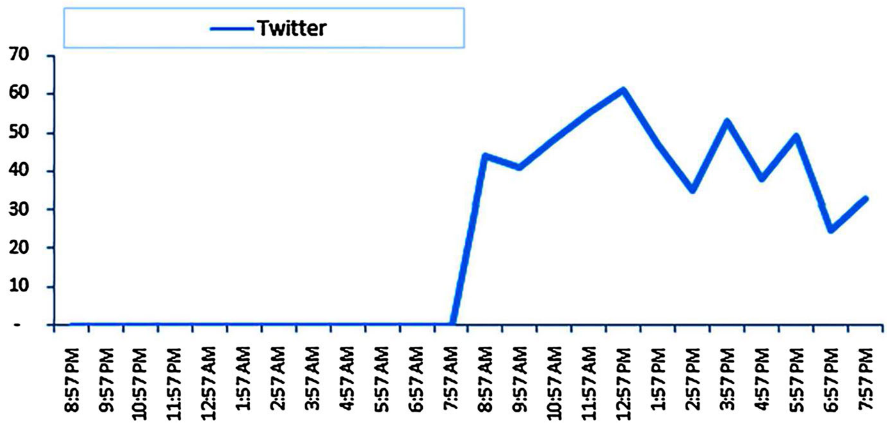


**Fig. 4** Snapshot of volume of tweets with respect to time for #sad

where, $P$ = # of positive instances, $N$ = # of negative instances, $TP$ = number of instances correctly labeled as belong to positive class, and $FP$ = items incorrectly labeled as belonging to the class.

- *Precision*
  The fraction of retrieved instances that are relevant is defined by:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

- *Recall*
  The fraction of relevant instances that are retrieved is defined by:

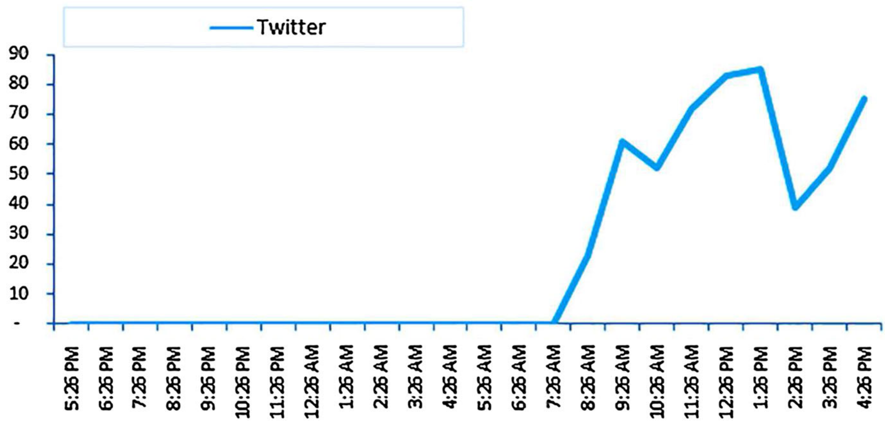


**Fig. 5** Snapshot of volume of tweets with respect to time for #surprise

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

where, $FN =$ items which were not labeled as belonging to the positive class but should have been.

- *F1 score*
  The F1 score (also known as balanced F-score or F-measure) of a system is the weighted harmonic mean of its precision and recall, and is being used to measure the accuracy of this test. It can have values between 1 (best) and 0 (worst). The corresponding formula for F1 Score is as follows:

$$F1\,score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

## 5 Findings and discussion

This section records and analyses the findings from the experiments conducted through unsupervised and supervised approaches (as discussed already).

### 5.1 Findings of unsupervised approach

We used SentiWordNet lexicon for the classification of tweets and movie review datasets. The proposed unsupervised methodology was applied to both datasets. Each tweet or movie review was scanned and every term would receive a score in view of SentiWordNet information and POS tagging. During the process, the NLTK POS tagger was used to find POS of the movie reviews and tweets, and Penn Treebank tags were used to access the score from SentiWordNet. A positive or negative weight was given to each tweet or review based on the scores received from SentiWordNet database. If the positive weight is greater than the negative weight, then we classify the tweet as a positive; otherwise, it is classified as a negative. We repeated this approach for the Polarity dataset and tweets retrieved from twitter public domain using keywords ‘Microsoft’, ‘Google’, ‘Twitter’, and ‘Apple’.

Using the model, as well as Google, to find the score for each term present in the adjective/noun/verb list, an accuracy of 80.68% was obtained for the tweets corpus. The result is shown in Table 2. It has been found that this approach not only addresses word sense disambiguation, but it is able to address another challenge in sentiment analysis, which is sudden deviation from positive sentiment to negative sentiment. A snapshot of the result for such a tweet, where there is a sudden change in polarity, is shown in Fig. 1.

### 5.2 Findings of supervised approach

We found that preparing a dataset by automatically collecting tweets using hashtags has an advantage as compared to the data set which is formed by manual annotation. This is instinctive, because writers are correct regarding their own emotions, while the conventional means of annotating information requires annotators to infer the authors’ emotions/feelings from the text, which can result in inaccurate classification.

- *For emotion dataset*
  People tend to use *hashtags* to express their sentiment or emotions, so these hash-tagged words are a good indication of sentiment and emotions. We used these hashtags to add more data to our machine learning algorithm. MNB classifier was used for the emotion dataset and the result is shown in Table 3. A snapshot of the confusion matrix of our emotion dataset for unigram features is shown in Fig. 6 and the F1 score of each class for unigram feature is shown in Table 4. ROC curve of the classifier is shown in Fig. 7.

- *For SMS dataset*
  The proposed model was evaluated using the collected SMS dataset, as SMS are more unstructured than tweets. Both MNB classifier and SVM were applied to the SMS data and the outcome is shown in Table 5. It was found that unigram

**Table 2** Comparison of performance between SentiWordNet Lexicon based approach and our proposed approach

| Approaches | Dataset | Accuracy (%) | Domain |
|---|---|---|---|
| SentiWordNet Lexicon | Reviews | 68.50 | Movie reviews |
| | Tweets | 75.20 | Google, Apple, Microsoft, Twitter |
| Proposed approach | Reviews | 69.0 | Movie reviews |
| | Tweets | 80.68 | Google, Apple, Microsoft, Twitter |

**Table 3** Accuracy of emotion dataset using different features

| Features | Number of features | MNB classifier (%) |
|---|---|---|
| Unigram | 4635 | 95.0 |
| Bigram | 17,628 | 71.23 |
| Unigram + Bigram | 35,356 | 95.3 |
| POS | 12,443 | 92.9 |
| Adjective | 1503 | 84.5 |

```
Confusion Matrix
             |                         t |
             |                         h |
             |                         a |
             |                         n |
             |                     s   k |
             |                     u   f |
             |                     r   u |
             |   a                 p   l |
             |   n   f       l     r   n |
             |   g   e   j   o   s i   e |
             |   e   a   o   v   a s   s |
             |   r   r   y   e   d e   s |
-------------+---------------------------+
       anger |<195>  2   .   2   .   .   . |
        fear |   .<207>  3   .   3   1   . |
         joy |   .   1<189>  .   6   .   . |
        love |   1   .   4<211>  4   .   . |
         sad |   .   1   .   2<179>  .   . |
    surprise |   .   1   .   .   4<203>  . |
thankfulness |   .   6   9   5  10   .  <.>|
-------------+---------------------------+
(row = reference; col = test)
```

**Fig. 6** Snapshot of emotion dataset

**Table 4** F1 score of MNB classifier for unigram feature

| Class label | Precision (%) | Recall (%) | F1 score (%) |
|---|---|---|---|
| Anger | 99.48 | 97.98 | 98.72 |
| Fear | 94.95 | 96.72 | 95.82 |
| Joy | 92.19 | 96.42 | 94.25 |
| Love | 95.90 | 95.90 | 95.9 |
| Sad | 86.89 | 98.35 | 92.26 |
| Surprise | 99.5 | 97.59 | 98.53 |

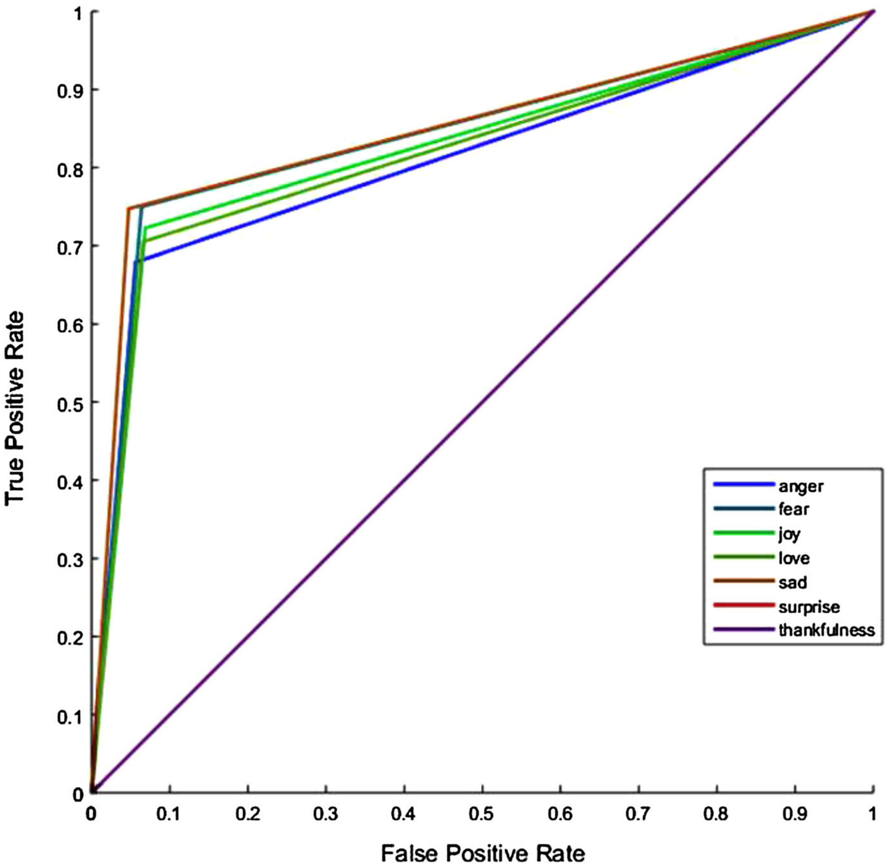


**Fig. 7** ROC curve of MNB classifier for emotion data set

**Table 5** Accuracy of SMS datset using different features

| Features | Number of features | MNB classifier (%) |
|---|---|---|
| Unigram | 2696 | 67 |
| Bigram | 21,041 | 63 |
| Unigram + Bigram | 42,008 | 60.2 |
| POS | 21,014 | 65 |
| Adjective | 1503 | 44 |

features are more useful for sentiment analysis of unstructured text. The results of using the unigram feature for the MNB classifier is reported in Table 6 and the corresponding ROC curve is shown in Fig. 8.

**Table 6** F1 score of MNB classifier for unigram feature

| Class label | Precision (%) | Recall (%) | F1 score (%) |
|---|---|---|---|
| Positive | 66.66 | 70.22 | 68.39 |
| Negative | 72.53 | 78.21 | 75.26 |
| Neutral | 62.5 | 20.83 | 31.24 |

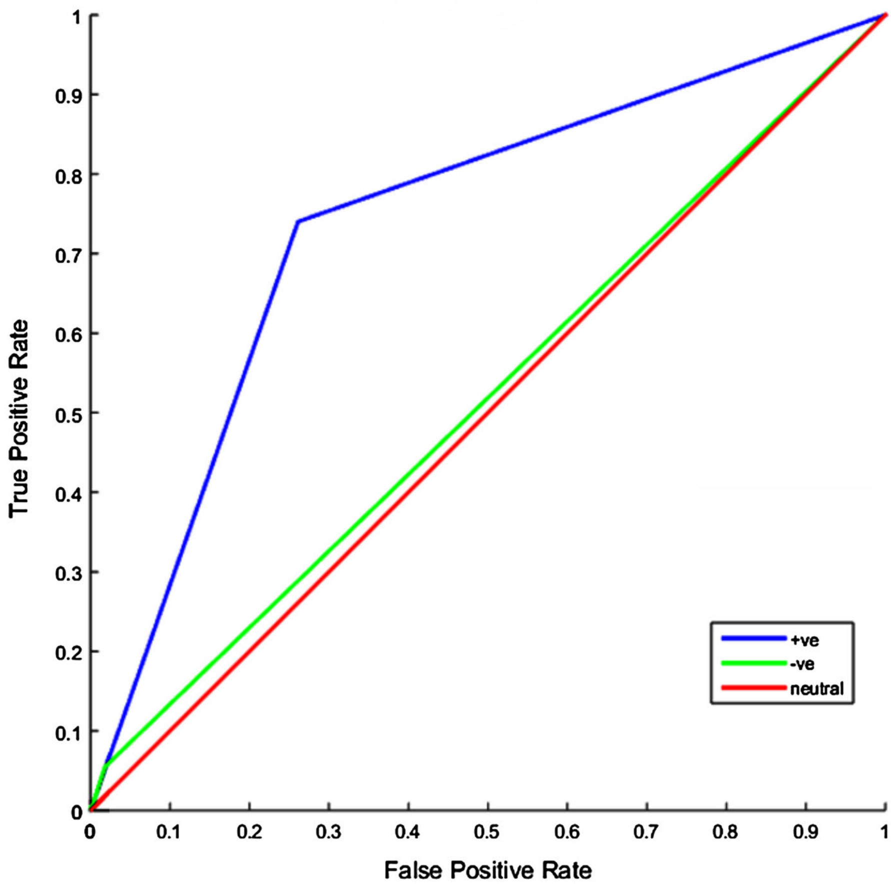


**Fig. 8** ROC curve of MNB classifier for SMS dataset

## 6 Conclusion

Sentiment analysis is a trend that is unlikely to fade due to the popularity of user-generated content on social media sites, etc. Many different techniques have been proposed in the literature to develop a reliable sentiment analysis system. However, due to the complexity and dynamic nature of social media data, it remains challenging to accurately identify sentiment in such data.

In this work, we evaluated the utility of unsupervised and supervised algorithms for sentiment classification of unstructured data such as tweets and SMS. We performed experiments on tweets to determine the effective features for sentiment analysis. While previous studies used SentiWordNet lexicon to build a sentiment analysis system for tweets, we observed that the accuracy obtained using SentiWordNet varies. Therefore, we generated a lexicon from our test corpus and used it for classification. One limitation with existing lexicons is that they may contain a large number of words with an associated sentiment score, and these lexicons lack words found in a particular domain. In comparison to the lexical resource, our proposed model leverages Google search engine to determine the score for each term using point wise mutual information. This results in a more robust dataset. It also allows us to address a particular challenge in sentiment analysis; that is, sudden deviation from positive polarity to negative polarity.

Machine learning algorithms were applied for identifying of sentiments in different datasets. We found that sentence level classification with features such as unigram presence and POS were most accurate, in comparison to other features used. However, for SMS (which are more unstructured than tweets), we suggested developing a dedicated system to identify term/word-specific sentiments in a SMS.

Future work includes extending this research to a larger dataset, with the aims of evaluating and refining our proposed approach, as well as exploring the potential to be deployed in review spam detection and intelligent recommender systems.

**Acknowledgements** This research is partially supported by the following projects: (1) Information Security Education & Awareness Project (Phase II), Ministry of Electronics and Information Technology (MeitY), Government of India, and (2) Grant No. ETI/359/2014 by Fund for Improvement of S&T Infrastructure in Universities and Higher Educational Institutions (FIST) Program 2016, Department of Science and Technology, Government of India.

## References

1. Bikel, D. M., & Sorensen, J. (2007). If we want your opinion. In: *International conference on semantic computing (ICSC 2007)* (pp. 493–500). doi:10.1109/ICSC.2007.81.
2. Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*(2), 15–21. doi:10.1109/MIS.2013.30.
3. Chen, R., & Xu, W. (2016). The determinants of online customer ratings: A combined domain ontology and topic text analytics approach. *Electronic Commerce Research*. doi:10.1007/s10660-016-9243-6.
4. Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In: *Proceedings of the 2008 international conference on web search and data mining* (pp. 231–240). doi:10.1145/1341531.1341561.
5. Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In: *Proceedings of 5th language resources and evaluation* (Vol. 6, pp. 417–422).
6. Fei, G., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2012). A dictionary-based approach to identifying aspects implied by adjectives for opinion mining. In: *Proceedings of 24th international conference on computational linguistics* (p. 309).
7. Feldman, R., Fresco, M., Goldenberg, J., Netzer, O., & Ungar, L. (2007). Extracting product comparisons from discussion boards. In: *Seventh IEEE international conference on data mining (ICDM 2007)* (pp. 469–474). doi:10.1109/ICDM.2007.27.

8. Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). Large-scale sentiment analysis for news and blogs. *Proceedings of the international conference on weblogs and social media (ICWSM)*, *7*(21), 219–222.
9. Hamouda, A., & Rohaim, M. (2011). Reviews classification using sentiwordnet lexicon. In: *World congress on computer science and information technology*.
10. Jindal, N., & Liu, B. (2006). Mining comparative sentences and relations. In: *Proceedings of the 21st national conference on artificial intelligence* (Vol. 2, pp. 1331–1336).
11. Van de Kauter, M., Breesch, D., & Hoste, V. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications*, *42*(11), 4999–5010. doi:10.1016/j.eswa.2015.02.007.
12. Li, J., Fong, S., Zhuang, Y., & Khoury, R. (2015). Hierarchical classification in text mining for sentiment analysis of online news. *Soft Computing*, *20*, 3411–3420. doi:10.1007/s00500-015-1812-4.
13. Li, Y., Qin, Z., Xu, W., & Guo, J. (2015). A holistic model of mining product aspects and associated sentiments from online reviews. *Multimedia Tools and Applications*, *74*(23), 10177–10194. doi:10.1007/s11042-014-2158-0.
14. Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, *2*, 627–666.
15. Liu, B. (2011). Opinion mining and sentiment analysis. In: *Web data mining: Exploring hyperlinks, contents, and usage data* (pp. 459–526). doi:10.1007/978-3-642-19460-3_11.
16. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. doi:10.2200/S00416ED1V01Y201204HLT016.
17. Liu, P., Gulla, J. A., & Zhang, L. (2016). Dynamic topic-based sentiment analysis of large-scale online news. In: *Proceedings of the 17th international conference on web information systems engineering* (pp. 3–18). doi:10.1007/978-3-319-48743-4_1.
18. Ma, Y., Chen, G., & Wei, Q. (2017). Finding users preferences from large-scale online reviews for personalized recommendation. *Electronic Commerce Research*, *17*(1), 3–29. doi:10.1007/s10660-016-9240-9.
19. Mo, S. Y. K., Liu, A., & Yang, S. Y. (2016). News sentiment to market impact and its feedback effect. *Environment Systems and Decisions*, *36*(2), 158–166. doi:10.1007/s10669-016-9590-9.
20. Montoyo, A., MartíNez-Barco, P., & Balahur, A. (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, *53*(4), 675–679. doi:10.1016/j.dss.2012.05.022.
21. Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2015). Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. *Expert Systems with Applications*, *42*(1), 306–324. doi:10.1016/j.eswa.2014.08.004.
22. Ohana, B. (2009). *Opinion mining with the sentwordnet lexical resource*. M.Sc. dissertation, Dublin Institute of Technology.
23. Pang, B., & Lee, L. (2004) A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of the 42nd annual meeting on association for computational linguistics* (p. 271). doi:10.3115/1218955.1218990.
24. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. doi:10.1561/1500000011.
25. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 conference on empirical methods in natural language processing* (Vol. 10, pp. 79–86). doi:10.3115/1118693.1118704.
26. Parkhe, V., & Biswas, B. (2016). Sentiment analysis of movie reviews: Finding most important movie aspects using driving factors. *Soft Computing*, *20*(9), 3373–3379. doi:10.1007/s00500-015-1779-1.
27. Peng, J., Choo, K. K. R., & Ashman, H. (2016). Astroturfing detection in social media: Using binary n-gram analysis for authorship attribution. In: *Proceedings of the 15th IEEE international conference on trust, security and privacy in computing and communications (TrustCom 2016)* (pp. 121–1286).
28. Peng, J., Choo, K. K. R., & Ashman, H. (2016). Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles. *Journal of Network and Computer Applications*, *70*, 171–182. doi:10.1016/j.jnca.2016.04.001.
29. Peng, J., Detchon, S., Choo, K. K. R., & Ashman, H. (2016). Astroturfing detection in social media: A binary n-gram-based approach. *Concurrency and Computation: Practice and Experience*. doi:10.1002/cpe.4013.

30. Pröllochs, N., Feuerriegel, S., & Neumann, D. (2015). Enhancing sentiment analysis of financial news by detecting negation scopes. In: *Proceedings of the 48th Hawaii international conference on system sciences (HICSS)* (pp. 959–968). doi:10.1109/HICSS.2015.119.
31. Robinson, R., Goh, T. T., & Zhang, R. (2012). Textual factors in online product reviews: A foundation for a more influential approach to opinion mining. *Electronic Commerce Research*, *12*(3), 301–330. doi:10.1007/s10660-012-9095-7.
32. Rout, J., Dalmia, A., Choo, K. K. R., Bakshi, S., & Jena, S. (2017). Revisiting semi-supervised learning for online deceptive review detection. *IEEE Access, 5*(1), 1319–1327. doi:10.1109/ACCESS.2017.2655032.
33. Rout, J., Singh, S., Jena, S., & Bakshi, S. (2017). Deceptive review detection using labeled and unlabeled data. *Multimedia Tools and Applications*, *76*(3), 3187–3211. doi:10.1007/s11042-016-3819-y.
34. Sadegh, M., Ibrahim, R., & Othman, Z. A. (2012). Opinion mining and sentiment analysis: A survey. *International Journal of Computers & Technology*, *2*(3), 171–178.
35. Song, L., Lau, R. Y. K., Kwok, R. C. W., Mirkovski, K., & Dou, W. (2017). Who are the spoilers in social media marketing? Incremental learning of latent semantics for social spam detection. *Electronic Commerce Research*, *17*(1), 51–81. doi:10.1007/s10660-016-9244-5.
36. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267–307. doi:10.1162/COLI_a_00049.
37. Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications*, *36*(7), 10760–10773. doi:10.1016/j.eswa.2009.02.063.
38. Turney, P.D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417–424). doi:10.3115/1073083.1073153.
39 Wang, D., Li, J., Xu, K., & Wu, Y. (2017). Sentiment community detection: Exploring sentiments and relationships in social networks. *Electronic Commerce Research*, *17*(1), 103–132. doi:10.1007/s10660-016-9233-8.
40 Zheng, L., Wang, H., & Gao, S. (2015). Sentimental feature selection for sentiment analysis of chinese online reviews. *International Journal of Machine Learning and Cybernetics*. doi:10.1007/s13042-015-0347-4.