Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble

Chapter · January 2018

DOI: 10.1007/978-3-319-74690-6_51

3 authors:

Mohammed Mamdouh Fouad
King Abdulaziz University

Tarek F Gharib
Ain Shams University

Abdulfatah S Mashat
King Abdulaziz University

Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble

Mohammed M. Fouad[1], Tarek F. Gharib[1(✉)], and Abdulfattah S. Mashat[2]

[1] Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
tfgharib@cis.asu.edu.eg

[2] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Abstract. Sentiment analysis from Twitter is one of the interesting research fields recently. It combines natural language processing techniques with the data mining approaches for building such systems. In this paper, we introduced an efficient system for Twitter sentiment analysis. The proposed system built a machine learning model for detecting positive and negative tweets. This model used different techniques to represent the input labeled tweets in the training phase using different features sets. In the classification phase, the classifier ensemble is presented with different base classifiers for more accurate results. The proposed system can be used for measuring users' opinion from their tweets which is very useful in many applications such as marketing, political polarity detection and reviewing products.

Keywords: Opinion mining · Sentiment analysis · Classifier ensemble
Feature selection · Information gain

1 Introduction

Nowadays, chatting, social media communications, blogging and micro-blogging, are the most utilized online activities on the Internet. Twitter is one of the most popular microblogging services and could be considered one of the largest user-generated content with very huge amount of structured and unstructured data. The posted tweets can express users' opinions about different topics and express their polarity towards these topics depending on the users' interests.

Sentiment analysis (also called opinion mining) uses a combination of data mining, text and web mining techniques in order to detect, extract and recognize the opinions, emotions and attitudes towards certain topics. It could be applied to different data sources such as review, blogs and news [1]. Detecting users' opinions is very useful information in many application domains such as evaluating marketing campaigns [2], reviewing movies [3], customer satisfaction [4] and many more.

Sentiment analysis in Twitter is different from other social media platforms due to many aspects: (i) users tend to use very short tweets to express their mood and status; (ii) users may use some abbreviations, emoticons to save up some characters; (iii) many linguistic representation challenges arise from feature engineering issues [5]. In this

A. E. Hassanien et al. (Eds.): AMLTA 2018, AISC 723, pp. 516–527, 2018.
https://doi.org/10.1007/978-3-319-74690-6_51

paper, we studied different combinations of features sets that could be used in tweets representation efficiently. As seen in most of the text mining systems, the extracted features may cause a complex computation problem due to the huge dimension of the generated features vector. To deal with such problem, features selection techniques such as information gain and mutual information [6], etc., or features transformation methods such as feature hashing [7] and principle component analysis [8] could be applied too.

Depending on the tweet message itself, it could be considered as a positive or a negative tweet if this message contains a sentiment with its text body, otherwise it is considered a neutral one. This led us to consider the sentiment analysis system as a classification problem. In such problem, we should analyze the input tweets collections and classify them with respect to the existing sentiments in each one. Moreover, the combination of multiple classifiers (called ensemble) is used to generate a single classifier to benefit from the properties of the individual classifiers. In this paper, we applied a majority voting ensemble that combines the decision from three base classifiers.

The main objective of this paper is to propose an efficient system for Twitter sentiment analysis using the information gain as a feature selection technique and the majority voting ensemble classifier. The proposed system is implemented and its accuracy is evaluated in order to answer the following **research questions**: (i) What is the prominent feature set that achieves the highest accuracy? And (ii) Does using information gain lead to better performance or not? And (iii) Does the ensemble model provide higher accuracy than the individual classifiers? Moreover, what are the factors that affect its performance?

The remainder of this paper is organized as follows: Sect. 2 contains the most relevant work in sentiment analysis problem. Section 3 represents in brief details the proposed system for twitter sentiment analysis. Experimental results and discussion are presented in Sect. 4. Conclusions are finally drawn in Sect. 5.

2 Related Work

Sentiment analysis is proposed to discover the users' polarity towards certain subject from their comments, reviews or opinions. This topic has been applied on news articles, blogs, product reviews, micro-blogs and forums. Due to the extensive research in this area, Ravi and Ravi [9] presented a detailed survey on the tasks, the approaches and the applications of the opinion mining that included a separate section for sentiment analysis in general. Another survey provided by Kharde and Sonawane [10] that covered the techniques of the sentiment analysis on Twitter data with comparative analysis of the existing approaches.

Ghiassi et al. [11] developed a Twitter-specific lexicon for sentiment analysis by utilizing a supervised feature selection technique using n-grams and statistical analysis. Their proposed model was tested using 3440 manually collected and annotated tweets from Justin Bieber Twitter account. Their experimental results show that their proposed model slightly outperformed the standard SVM classifier and achieved 95.1% accuracy.

Selecting the prominent features set for sentiment analysis is one of the challenges with the existing sentiment analysis methods. There are several types of features that could be extracted from the tweets text, but what are the combination that achieves the highest accuracy rate. Recently, Agrawal and Mittal [12] explored various feature extraction and selection techniques to discover the prominent features in a machine learning based sentiment analysis. They combined the lexicon-based approaches with the corpus-based approaches to find the semantic orientation of all the extracted features to measure the overall polarity of the input text.

Agrawal et al. [13] studied the feature-engineering problem on twitter sentiment classification. Their feature sets combined different features such as unigram, POS-features, senti-features and tree kernel model. For the classification task, SVM was used with the different feature sets combinations. They applied their proposed system to a collection of 11,875 tweets that were manually annotated. According to their results, the feature set containing the unigram and senti-features achieved the highest accuracy rate with about 75.39%.

Most of the presented machine learning based sentiment analysis methods used a single classifier to perform the classification task. For example, Zhang et al. [14] and Mohammad et al. [15] used Support Vector Machines (SVM) algorithm, while others like Saif et al. [16] utilized the Naïve Bayes (NB) algorithm because of their good performance in text classification problems. On the other hand, classifier ensemble approach is introduced to train multiple classifiers and combine their decisions to solve the same classification problem. This approach tried to cover some of the problems of the individual classifiers by combining different classifiers tending to produce a generalized decision boundary for the classification input [17]. It is not guaranteed that the performance of the classifier ensemble is always better than the individual classifiers combined in it, but in some cases, it reduces the risk of selecting inefficient classifier with the unseen data [18].

The proposed classifier ensembles are different in the base classifiers used in each ensemble and the way it combines their decisions. For example, both Lin and Koltz [19] and Rodríguez-Penago et al. [20] used the majority voting ensemble in their work. Clark et al. [21] used the weighted voting ensemble with trained Naïve Bayes classifiers. In addition, Hassan et al. [5] proposed a bootstrap model that combined different dataset, feature and classifier parameters with utilization of about 6 base classifiers. Recently, Da Silva et al. [22] presented another combination rule. They calculated the average of the probabilities that were produced by four classifiers for each class as the final decision of the ensemble classifier.

3 The Proposed System

In this section, we will describe in brief details the components of the proposed system for twitter sentiment analysis. As shown in Fig. 1, the proposed system is running in two phases: Training and Classification phases. The purpose of the training phase is to build the classification model in order to distinguish between positive and negative tweets based on the input labeled tweets collections. In the classification phase, the trained classification model will assign positive or negative label to the new unlabeled tweets. The system contains four steps: Preprocessing, Feature Extraction, Feature Selection and the Classification Model for Sentiment Analysis.

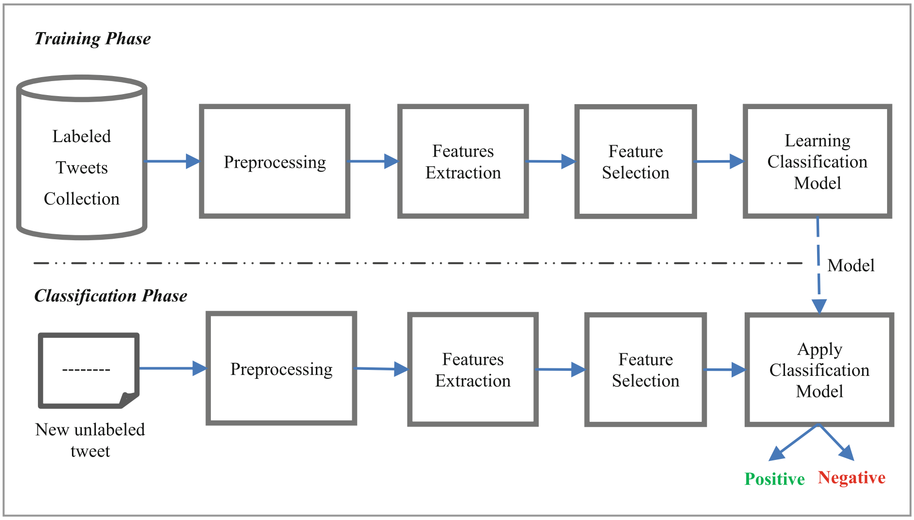

Fig. 1. Overview of the proposed system

3.1 Preprocessing

The main objective of this step is to use the natural language processing techniques to process the input tweet text and make it suitable for the next step to extract the features correctly. The detailed block diagram with example tweet for the preprocessing step is shown in Fig. 2.

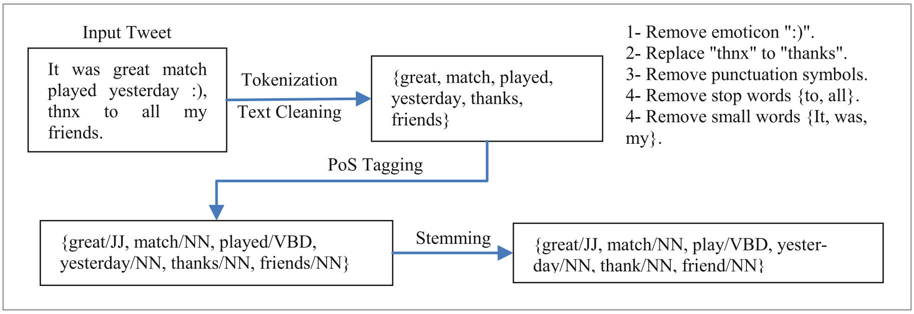

Fig. 2. Example for tweet text preprocessing

The preprocessing step includes four sub-steps: Tokenization, Text Cleaning, PoS Tagging and Stemming. The preprocessing started with tokenizing (splitting) the input text into separate terms (called tokens). Each token can represent word, abbreviation, hyperlink, emoticon or other punctuation symbols that could be found commonly in tweets.

The second step is the Text Cleaning step, which is responsible for removing any irrelevant textual data from the tweet content itself. As shown in Fig. 2, the example input tweet "It was great match played yesterday :), thnx to all my friends" is transformed into a list of words which is {great, match, played, yesterday, thanks, friends} after applying the tokenization and text cleaning steps.

The third step is PoS Tagging in which we extract the part of speech tags for the input text. For example, word such as "great" is tagged with "JJ" because it is an adjective word. The final step is to stem the words to their original root in order to reduce the initial set of words representing the input tweet text. For example, the word "played" is transformed into the stem word "play". The final words of the preprocessing step for the example tweet is {great/JJ, match/NN, play/VBD, yesterday/NN, thank/NN, friend/NN}.

3.2 Feature Extraction

There are several features to be extracted to represent the input tweet text. In this section, we will present the different types of features that will be used in the proposed system.

The most simple and traditional technique in text representation is to extract all the possible stemmed words, may also called terms or tokens, from the input text which is called **Bag-of-words (BoW)**. In this paper, the BoW contains all the distinct unigrams (single word terms) and bigrams (two consecutive words terms). For example tweet "great match play yesterday thank friend", the unigram features are "great", "match", "play", "yesterday", "thank" and "friend". The bigram features are "great_match", "match_play", "play_yesterday", "yesterday_thank" and "thank_friend".

Some words can express the opinion state of the writer. Words like *great*, *good*, *wonderful* and *excellent* can express positive opinion, while words such as *poor*, *bad* and *dangerous* are examples of negative opinion. In this paper, we used the opinion lexicon collected by Liu et al. [23] that contains a list of 2006 positive words and 4783 negative words. For each tweet, the positive and negative words are counted as the **Lexicon-based features**. For the example tweet, it has *two* positive words, "*great*" and "*thank*", and *zero* negative words as lexicon-based features.

During the preprocessing step, the part-of-speech tags for the extracted words are stored. We count the numbers of nouns, verbs, adjectives and adverbs as the **PoS features**. For the example tweet, the extracted PoS features are four nouns ("match", "yesterday", "thank" and "friend"), one verb ("play"), one adjective ("great") and zero adverbs.

Emoticons are some symbols that represent certain state to the writer opinion. In this step, we collected a list of commonly used emoticons used in the social media and especially in tweets. The list contains 112 positive, 77 negative and 16 neutral emoticon symbols. For each tweet in the collection, the number of the found emoticons in each state is recorded as the **Emoticons features**. For the example tweet, it has only *one* positive emoticon, which is :), and *zero* negative and neutral ones.

3.3 Feature Selection

As discussed earlier, each tweet is represented by a vector of numbers based on the extracted features. The biggest portion of these features follows to the BoW unigrams and bigrams. The dimension of this vector increased dramatically by the number of distinct terms in the input tweets collection. The curse of high dimensionality exists in most of the text processing systems including sentiment analysis ones. For this case, we used Information Gain (IG) as feature selection technique to reduce the dimension of the output feature vector. In the proposed system, the information gain weight is calculated for each feature using Eq. (1) and the features that have higher weight than 0.01 are selected.

Consider the input tweets collection with class attribute C that has two classes $\{C_1 = \text{positive and } C_2 = \text{negative}\}$. For any given feature x, the information gain (IG) is calculated by:

$$IG(x) = -\sum_{j=1}^{2} P(C_j)\log\left(P(C_j)\right) + P(x)\sum_{j=1}^{2} P(C_j|x)\log\left(P(C_j|x)\right) + P(\bar{x})\sum_{j=1}^{2} P(C_j|\bar{x})\log\left(P(C_j|\bar{x})\right) \tag{1}$$

Where, $P(C_j)$ is the fraction of tweets labeled with class C_j, $P(x)$ is the fraction of tweets in which feature x occurs and $P(C_j|x)$ is the fraction of tweets with class C_j that has feature x.

3.4 Classification Model for Sentiment Analysis

The main step in the proposed system is to build a classification model that is able to differentiate efficiently between positive and negative labeled tweets in the training phase.

There are several machine learning algorithms that could be used in building such a model. In the proposed system, we implemented a majority voting ensemble classifier with SVM, NB and LR as base learners. These algorithms are commonly used and have great success in the text classification problems. An overview of the majority voting classifier ensemble used in the proposed system is shown in Fig. 3.

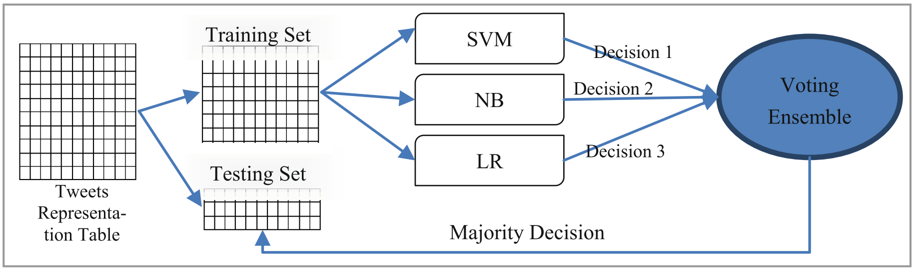

Fig. 3. Majority voting classifier ensemble

As shown in Fig. 3, the extracted features from the input tweets collection are split into two sets, training and testing set. The training set is passed to each classifier to register its decision. Then the final decision output from the voting ensemble is the majority decision obtained from the three classifiers. The voting ensemble model will consider this tweet as positive one because this is the majority decision. Finally, the testing set is used to validate the accuracy of the built classifier ensemble model.

4 Experimental Results and Discussion

The proposed system is implemented and its accuracy is measured against different well-known datasets in the area of twitter sentiment analysis. The preprocessing and feature extraction steps are implemented in Java with the support of Stanford CoreNLP library. The feature selection and the classification ensemble model are implemented using RapidMiner® tool. All the experiments were conducted using a machine with Intel® Core™ i7-3770 CPU @ 3.40 GHz and 8.00 GB memory, running 64-bit Windows 7 Enterprise Edition®.

4.1 Datasets

Four datasets are used in evaluating the designed experiments in order to evaluate the performance of the proposed system. The distribution of the positive and negative polarities in the used datasets is shown in Table 1.

Table 1. Distribution of positive and negative polarities in the datasets

Dataset	Number of tweets	
	Positive	Negative
Stanford-1K	500	500
Stanford-3K	1500	1500
Sanders	201	293
HCR	211	386

Stanford Twitter Sentiment Corpus contains about 1.6M tweets (800,000 positive and 800,000 negative tweets) collected by a scrapper that calls Twitter API for some queries [24]. In our experiments, we did not use the complete training dataset due to the computational limitations. We perform unified sampling to obtain two sample datasets, Stanford-1K and Stanford-3K with 1000 and 3000 tweets respectively.

The third data sets is called **Sanders Dataset** [25] which contains about 5513 hand classified tweets with four labels: positive, negative, neutral and irrelevant. Twitter API is used with four search terms: @apple, #google, #microsoft and #twitter. We could not obtain all these tweets because most of them are currently invalid or deleted. In our experiments, we are interested in positive and negative labeled tweets only, which are about 201 positive and 293 negative tweets.

The fourth dataset is called **Health Care Reform (HCR) Dataset**. This dataset is collected by crawling tweets that contain the hashtag "#hcr" in March 2010 [26]. Some of these tweets are manually labeled into positive, negative and neutral label. In our experiment, we are interested in positive and negative ones only, which are about 597 tweets (211 positive and 386 negative).

4.2 Sentiment Analysis Results

Prominent Features Set. The objective of this experiment is to answer the first research question and decide what are the prominent features set that achieves the highest accuracy. Table 2 reports the accuracy of the proposed system using the majority voting ensemble model. Each dataset is represented with the different combinations of the Bag-of-Words (BoW), Lexicon-based features (Lex), Emoticon-based features (Emo) and PoS features (PoS).

Table 2. Accuracy for different combinations of features sets

Features Set	Accuracy (%)			
	Stanford-1K	Stanford-3K	Sanders	HCR
BoW	73.90	76.00	93.53	84.58
BoW + Lex	77.90	76.53	93.73	83.91
BoW + Emo	74.50	75.27	93.33	**84.75**
BoW + PoS	74.00	75.60	93.53	84.41
BoW + Lex + Emo	**78.70**	76.57	93.73	**84.75**
BoW + Lex + PoS	77.30	**77.27**	**93.94**	84.41
BoW + Emo + PoS	74.40	75.63	93.34	84.58
BoW + Lex + Emo + PoS	**78.70**	77.00	93.53	**84.75**

As shown in Table 2, the difference in the accuracy between BoW and other different combinations of features sets is not huge. Also, we notice that using the feature set that includes all the features (BoW + Lex + Emo + PoS) leads to the better accuracy as in Stanford-1K and HCR datasets. In Stanford-3K and Sanders datasets, the accuracy of the proposed system when using the complete features set still very high with very small margin compared to the highest accuracy. This allows us to state that *Lex and PoS features are good additions to traditional BoW features while Emo features do not add much to the overall accuracy.*

Using Information Gain (IG). The aim of this experiment is to answer the second research question and examine the effect of using information gain technique on the proposed system in two aspects: the accuracy enhancement and the dimension reduction. In the first part of the experiment, the accuracy of the standalone classifiers (SVM, LR and NB) and the Majority Voting Ensemble (MVE) model are reported in two cases, without using the information gain and with it applied.

In the second part of the experiment, we are interested in measuring the reduction ratio obtained after using information gain technique, because IG technique is mainly used to select the features that better match the given classes. Table 3 shows the comparison of the reported accuracy for each dataset and the feature vector length before and after using the IG technique.

Table 3. Accuracy comparison after using IG with different classifiers

Dataset		Accuracy (%)				Feature vector length	
		SVM	LR	NB	MVE	# Features	Reduction (%)
Stanford-1K	Without IG	63.6	65.5	62.4	64.8	911	**61.14**
	With IG	**78.1**	**74.5**	**76.5**	**78.7**	557	
Stanford-3K	Without IG	66.73	63.33	61.37	65.37	2400	**46.75**
	With IG	**79.1**	**71.13**	**77.77**	**77.27**	1122	
Sanders	Without IG	80.77	79.57	79.35	81.79	1023	**76.44**
	With IG	**92.71**	**90.11**	**91.91**	**93.94**	782	
HCR	Without IG	72.7	65.17	67.85	69.86	1357	**69.49**
	With IG	**81.22**	**75.37**	**85.09**	**84.75**	943	

As shown in Table 3, it is clear that using information gain technique enhanced the accuracy of the proposed system in all used classifiers with about 15% on average. In addition, using IG technique reduces the feature vector length for each dataset with about 63.45% on average. This allows the classifiers to distinguish efficiently between positive and negative classes with lower computational requirements. From these results, we can conclude that using the information gain (IG) technique not only reduces the dimension of the feature vector greatly, but also enhances the performance of the model classifier very efficiently.

Majority Voting Ensemble (MVE) Evaluation. In this experiment, we target to evaluate the performance of the MVE model in order to answer the third research question. The performance of each classifier is measured for all the datasets and the parameters that achieved the highest accuracy are recorded. MVE model is also tested by combining the decisions of the standalone classifiers with the optimal parameters and its accuracy is also recorded to be compared with other algorithms as shown in Table 4.

Table 4. Accuracy comparison for different classifiers

Classifier	Best accuracy (%)			
	Stanford-1K	Stanford-3K	Sanders	HCR
SVM	78.10	**79.10**	92.71	81.22
LR	74.50	71.13	90.11	75.37
NB	76.50	77.77	91.91	**85.09**
MVE	**78.70**	77.27	**93.94**	84.75

As shown in Table 4, SVM and NB classifiers have good performance with respect to different datasets, while LR has the worst accuracy results. Regarding the Majority Voting Ensemble (MVE) model, its performance is affected by the performance of the individual classifiers. For example, in Stanford-1K and Sanders datasets, when the base classifiers have good results, MVE outperforms them and achieve better results with accuracy 78.70% and 93.94% respectively. In HCR dataset, LR achieves very low accuracy, about 75.37%, compared to SVM and NB results. We can see that MVE tries to recover from such drop and achieves about 84.75% accuracy with a little bit difference to the highest accuracy that achieved by NB classifier (about 85.09%).

From these results, we can notice that MVE achieves the highest accuracy, and even outperforms the individual classifiers, when the results of these classifiers are near. In the case that one classifier has soft performance; MVE tries to recover from such performance and achieves good results which are very near to the best ones.

5 Conclusions and Future Work

In this paper, we introduced an efficient system for Twitter sentiment analysis. The proposed system used different techniques to represent input labeled tweets with different features sets. The irrelevant and insignificant features are early pruned using Information Gain (IG) feature selection technique. The classifier ensemble is built from diversified set of base classifier to perform the classification task which is responsible for detecting the output sentiment polarity. Many experiments were conducted using the most commonly used tweets datasets to analyze the performance of the proposed system in different aspects.

The experimental results answered the three main research questions in this work. First, using IG feature selection technique boosted the accuracy of the individual classifiers and the ensemble model with about 15% on average. Second, the ensemble model tried to combine the performance of the base classifiers, but its results could be affected if one classifier was not suitable for the used dataset. Third, we can notice that the reported results of the lexicon-based features and PoS features enhanced the accuracy of the classifiers when added to the BoW features. On the other hand, emoticon-based features did not have this much addition.

As a future work, we may include the "neutral" tweets into the proposed system by adapting the feature extraction and classification steps to recognize these tweets efficiently. In addition, to support tweets from other languages, such as Arabic [27], the proposed system could be adapted to be multi-lingual system.

References

1. Gaur, M., Pruthi, J.: A survey on sentiment analysis and opinion mining. Int. J. Curr. Eng. Technol. **7**(2), 444–446 (2017)
2. Li, Y.-M., Li, T.-Y.: Deriving market intelligence from microblogs. Decis. Support Syst. **55**(1), 206–217 (2013)
3. Rui, H., Liu, Y., Whinston, A.: Whose and what chatter matters? The effect of tweets on movie sales. Decis. Support Syst. **55**(4), 863–870 (2013)
4. Kang, D., Park, Y.: Review-based measurement of customer satisfaction in mobile service: sentiment analysis and VIKOR approach. Expert Syst. Appl. **41**(4), 1041–1050 (2014)

5. Hassan, A., Abbasi, A., Zeng, D.: Twitter sentiment analysis: a bootstrap ensemble framework. In: The International Conference on Social Computing (SocialCom), Alexandria, VA (2013)
6. Manning, C., Raghvan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
7. Lin, J., Kolcz, A.: Large-scale machine learning at twitter. In: The International Conference on Management of Data (SIGMOD 2012), New York, NY, USA (2012)
8. Vinodhini, G., Chandrasekaran, R.: Sentiment classification using principal component analysis based neural network model. In: The International Conference on Information Communication and Embedded Systems (ICICES 2014), Chennai, India (2014)
9. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl. Based Syst. **89**, 14–46 (2015)
10. Kharde, V., Sonawane, S.: Sentiment analysis of twitter data: a survey of techniques. Int. J. Comput. Appl. **139**(11), 5–15 (2016)
11. Ghiassi, M., Skinner, J., Zimbra, D.: Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network. Expert Syst. Appl. **40**, 6266–6282 (2013)
12. Agarwal, B., Mittal, N.: Prominent Feature Extraction for Sentiment Analysis. Socio-Affective Computing Series. Springer International Publishing (2016)
13. Agrawal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, P.: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Languages in Social Media (LSM 2011), Stroudsburg, PA, USA (2011)
14. Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B.: Combining lexicon-based and learning-based methods for twitter sentiment analysis. HP Laboratories (2011)
15. Mohammad, S., Kiritchenko, S., Zhu, X.: NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of the Seventh International Workshop on Semantic Evaluation Exercises (SemEval-2013), Atlanta, GA, USA (2013)
16. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of twitter. In: Proceedings of the 11th International Conference on the Semantic Web (ISWC 2012), Berlin, Heidelberg (2012)
17. Kuncheva, L.: Combining Pattern Classifiers: Methods and Algorithms, 2nd edn. Wiley, New York (2014)
18. Fumera, G., Roli, F.: A theoretical and experimental analysis of linear combiners for multiple classifier systems. IEEE Trans. Pattern Anal. Mach. Intell. **27**(6), 942–956 (2005)
19. Lin, J., Kolsz, A.: Large-scale machine learning at twitter. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (SIGMOD 2012), New York, NY, USA (2012)
20. Rodríguez-Penagos, C., Atserias, J., Codina-Filba, J., Garcıa-Narbona, D., Grivolla, J., Lambert, P., Saur, R.: FBM: combining lexicon-based ML and heuristics for social media polarities. In: Proceedings of the Seventh International Workshop on Semantic Evaluation, Atlanta, GA, USA (2013)
21. Clark, S., Wicentwoski, R.: SwatCS: combining simple classifiers with estimated accuracy. In: Proceedings of the Seventh International Workshop on Semantic Evaluation, Atlanta, GA, USA (2013)
22. Da Silva, N., Hruschka, E., Hruschka Jr., E.: Tweet sentiment analysis with classifier ensembles. Decis. Support Syst. **66**, 170–179 (2014)
23. Liu, B., Hu, M., Cheng, J.: Opinion observer: analyzing and comparing opinions on the web. In: 14th International World Wide Web Conference, Chiba, Japan (2005)
24. Stanford Twitter Sentiment Corpus. http://help.sentiment140.com/for-students. Accessed May 2017

25. Sanders Dataset. http://www.sananalytics.com/lab/. Accessed May 2017
26. Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.: Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of the First Workshop on Unsupervised Learning in NLP (EMNLP 2011), Stroudsburg, PA, USA (2011)
27. Mostafa, A.M.: An evaluation of sentiment analysis and classification algorithms for Arabic textual data. Int. J. Comput. Appl. **158**(2), 29–36 (2017)