

# Tweet Sentiment Classification Using an Ensemble of Machine Learning Supervised Classifiers Employing Statistical Feature Selection Methods

K. Lakshmi Devi, P. Subathra and P.N. Kumar

**Abstract** Twitter is considered to be the most powerful tool of information dissemination among the micro-blogging websites. Everyday large user generated contents are being posted in Twitter and determining the sentiment of these contents can be useful to individuals, business companies, government organisations etc. Many Machine Learning approaches are being investigated for years and there is no consensus as to which method is most suitable for any particular application. Recent research has revealed the potential of ensemble learners to provide improved accuracy in sentiment classification. In this work, we conducted a performance comparison of ensemble learners like Bagging and Boosting with the baseline methods like Support Vector Machines, Naive Bayes and Maximum Entropy classifiers. As against the traditional method of using Bag of Words for feature selection, we have incorporated statistical methods of feature selection like Point wise Mutual Information and Chi-square methods, which resulted in improved accuracy. We performed the evaluation using Twitter dataset and the empirical results revealed that ensemble methods provided more accurate results than baseline classifiers.

**Keywords** Bagging · Boosting · Ensemble learners · Entropy · Naïve bayes · Sentiment classification · SVM

---

K.L. Devi · P. Subathra · P.N. Kumar (✉)  
Department of CSE, Amrita Vishwa Vidyapeetham, Coimbatore 641112, India  
e-mail: pn\_kumar@cb.amrita.edu

K.L. Devi  
e-mail: laksdevi115@gmail.com

P. Subathra  
e-mail: p\_subathra@cb.amrita.edu

## 1 Introduction

Micro-blogging websites have become valuable source of information which varies from personal expressions to public opinions. People post status messages about their life, share opinions on their political and religious views, express support or protest against social issues, discuss about products in market, education, entertainment etc. Hence micro-blogging websites have become a powerful platform for millions of users to express their opinion. Among all micro-blogging websites, the most popular and powerful tool of information dissemination is Twitter launched in October 2006, which allows users to post free format textual messages called tweets. Owing to the free format of messages, users are relieved from concerning about grammar and spelling of the language. Hence Twitter allows short and fast posting of tweets where each tweet is of length 140 characters or less. These tweets are indicative of views, attitudes and traits of users and therefore they are rich sources of sentiments expressed by the people. Detecting those sentiments and characterizing them are very important for the users and product manufacturers to make informed decisions on products. Hence Twitter can be labeled as a powerful marketing tool. Recently, there has been a shift from blogging to micro-blogging. The reason for this shift is that micro-blogging provides faster mode of information dissemination when compared to blogging. Since there is restriction on the number of characters allowed, it reduces users' time consumption and effort for content generation. Another major difference is that the frequency of updates in micro blogging is very high when compared to blogging due to shorter posts [1].

Sentiment classification of social media data can be performed using many Machine Learning (ML) algorithms among which Support Vector Machine (SVM), Naive Bayes (NB), Maximum Entropy (Maxent) are widely used. These baseline methods use Bag Of Words (BOW) representation of words, where frequency of each word is used as a 'feature' which is fed to the classifier. These models ignore the word order but maintain the multiplicity. Ensemble techniques have gained huge prominence in the Natural Language Processing (NLP) field and they have been proved to provide improved accuracy. This is achieved by combining different classifiers which are trained using different subsets of data, resulting into a network of classifiers which can be further used to perform sentiment classification [2]. The major limitation associated with ensemble learners is that the training time is very huge since each of the classifiers must be individually trained. This increases the computational complexity as the dimensionality of the data increases. Therefore, the ensemble learning techniques are employed where the data has fewer number of dimensions and also when maximum possible accuracy of classification is a mandatory requirement [2].

In this work, we mainly focus on the sentiment classification of tweets using ensemble learners like Bagging and Boosting and performed a comparative analysis with base learners viz. SVM, NB and Maxent classifiers. We have employed Twitter dataset in this work and the results demonstrate that the ensemble methods outperformed baseline methods in terms of accuracy. We have adopted statistical

feature selection methods like Point wise Mutual Information (PMI) and Chi-square methods for selecting the most informative features which are then fed to the classifier models. The rest of the paper is organized as follows: Sect. 2 discusses the related work. Section 3 describes the proposed system, feature selection methods, baseline methods and ensemble learners. Section 4 describes the implementation details and experiments. The results and analysis is given in Sect. 5. The conclusion and future enhancements of the paper are given in Sect. 6.

## 2 Related Work

Sentiment Analysis (SA) can be broadly classified into three: Sentence level SA, Document level SA and Aspect level SA [3]. The document level SA considers a single whole document as the fundamental unit assuming that the entire document talks about a single topic. The sentence level SA can be considered as a slight variation of document level SA where each sentence can be taken as a short document [4]. Each sentence can be subjective or objective and SA is performed on the subjective sentences to determine the polarity i.e., positive or negative. Aspect level SA [5] is performed when we have to consider different aspects of an entity. This is mostly applicable in the SA of product reviews, for example: “The battery life of the phone is too low but the camera quality is good”. SA can be performed in other levels of granularity like clause, phrase and word level depending on the application under consideration. The sentiment classification approaches can be divided into three

- Machine Learning (ML) approach
- Lexicon Based approach
- Hybrid approach.

### 2.1 *Machine Learning (ML) Approach*

In machine learning approach, different learning algorithms like NB, SVM, Maxent etc., use linguistic features to perform the classification. The ML approaches can be supervised, where a labeled set of training data is provided, or unsupervised which is used where it is difficult to provide labeled training data. The unsupervised approach builds a sentiment lexicon in an unsupervised way and then evaluates the polarity of text using a function that involves positive, negative and neutral indicators [6]. The major goal aim of a semi-supervised learning approach is to produce more accurate results by using both labeled and unlabeled data [7]. Even though there exist unsupervised and semi-supervised techniques for sentiment classification, supervised techniques are considered to have more predictive power [6].

## ***2.2 Lexicon Based Approach***

This approach is based on the already available pre compiled collection of sentiment terms which is called as ‘sentiment lexicon’. This approach aims at discovering the opinion lexicon and is classified into two: the dictionary based approach and corpus based approach. The dictionary-based approach works by considering some seed words which is a predefined dictionary that contains positive and negative words for which the polarity is already known. A dictionary search is then performed in order to pick up the synonyms and antonyms of these words and it then adopts word frequency or other measures in order to score all the opinions in the text data. The computational cost for performing automated sentiment analysis is very low when the dictionary is completely predefined [8]. The corpus based approach builds a seed list containing opinion words and uses semantics and statistics to find other opinion words from a huge corpus in order to find opinion words with respect to the underlying context [6].

## ***2.3 Hybrid Approach***

The third approach for performing sentiment classification is the hybrid approach which is a combination of machine learning and lexicon based approach. The hybrid methods have gained huge prominence in recent years. This approach aims at attaining the best of both approaches i.e. robustness and readability from a well-designed lexicon resource and improved accuracy from machine learning algorithms. The main aim of these approaches is to classify the text according to polarity with maximum attainable accuracy.

## ***2.4 Ensemble Methods***

Each of these approaches work well in different domains and there is no consensus as to which approach performs well in a particular domain. So, in order to mitigate these difficulties, ‘an ensemble of many classifiers’ can be adopted for achieving much more accurate and promising results. Recent research has revealed the potential of ensemble learners to provide improved accuracy in sentiment classification [9]. An ensemble should satisfy two important conditions: prediction diversity and accuracy [6]. Ensemble learning comprises of state of the art techniques like Bagging [10], Boosting [11] and Majority Voting [12]. Majority voting is the most prominent ensemble technique used in which there exists a set of experts which can classify a sentence and identify the polarity of the sentence by choosing the majority label prediction. This results in improved accuracy; however, this method does not address prediction diversity. Bagging and Boosting techniques are

explicitly used to address the issue of prediction diversity. The bagging technique works by random sampling with replacement of training data i.e. every training subset drawn from the entire data (also called as bag) is provided to different baseline methods of learner of the same kind [9]. These randomly drawn subsets are known as bootstrapped replicas of entire training data. Boosting techniques generally constructs an ensemble incrementally in which a new model is trained with an emphasis on the instances which were misclassified by the previous models [10].

### 3 Proposed Work

In this paper, we have implemented Sentiment multi-class Classification of twitter dataset using Ensemble methods viz. Bagging and Boosting. A comparative analysis with baseline methods SVM, NB and Maxent has been done and their performance evaluated using evaluation metrics accuracy, precision, recall and f-measure. We have also employed statistical feature selection methods like PMI and Chi-square to generate n most informative features. The proposed system consists of:

- Tweet pre-processing.
- Feature selection using Statistical methods like PMI and Chi-square.
- Multi class Sentiment Classification into positive, negative and neutral.
- Performance Evaluation.

The system architecture is given in Fig. 1.

#### 3.1 *Tweet Pre-processing*

Since Twitter allows free format of messages, people generally do not care about grammar and spelling. Hence the dataset must be pre-processed before it can be used for sentiment classification. The pre-processing steps include:

- Tokenizing.
- Removing non-English tweets, URLs, target (denoted using @), special characters and punctuations from hash tags (a hash tag is also called as summarizer of a tweet), numbers and stop words.
- Replacing negative mentions (words that end with 'nt' are replaced with 'not'), sequence of repeated characters (For example, 'wooooooww' is replaced by 'wooooww').

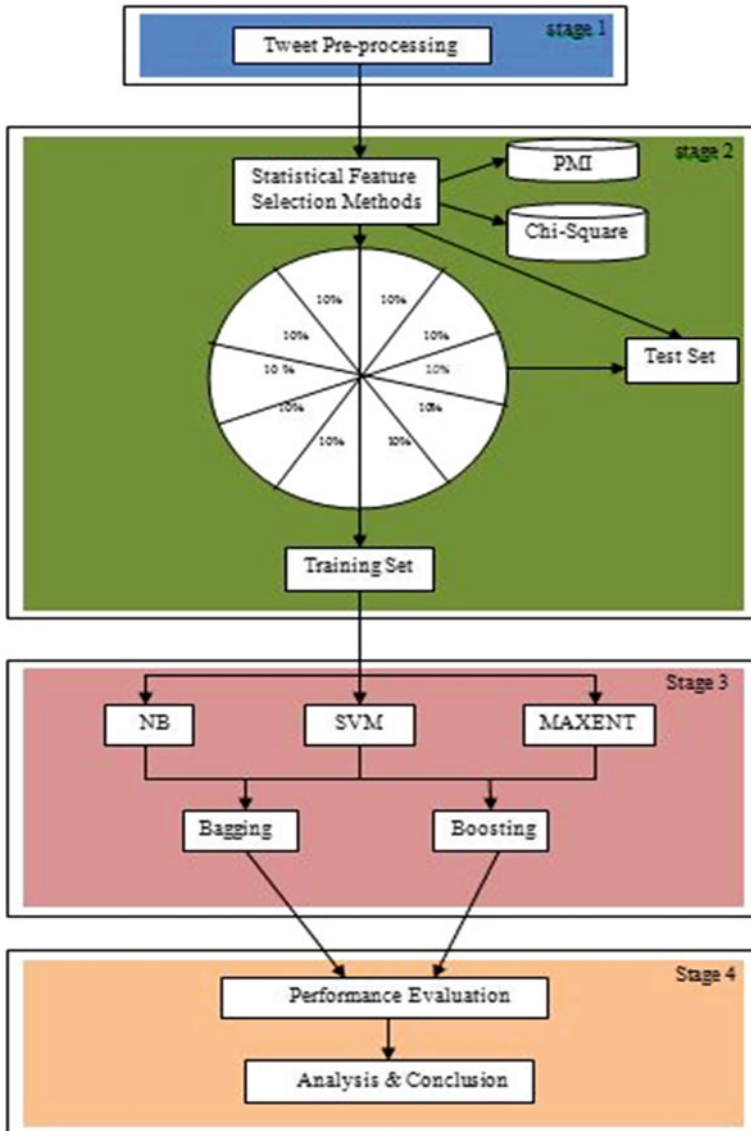


Fig. 1 Architecture diagram of sentiment multi-class classification

### 3.2 Feature Selection

Feature selection methods treat each sentence/document as a BOW or as a string that maintains the sequence of the words that constitute the sentence/document. BOW which is known for its simplicity is widely used by baseline methods for

sentiment classification which often adopts stop word elimination, stemming etc. for feature selection. The feature selection methods can be broadly classified into two:

- Lexicon based: This method incrementally builds large lexicon using a set of ‘seed words’ and gradually bootstrapping the set by identifying synonyms and incorporating available online resources. The main limitation associated with this approach is that this method needs human annotation.
- Statistical method: This approach is fully automatic and they are widely adopted in sentiment analysis. We have employed two statistical methods of feature selection in this work.

### 3.3 Point Wise Mutual Information (PMI)

The mutual information measure indicates how much information one particular word provides about the other. In text classification, this method can be used to model the mutual information between the features selected from the text and the classes. The PMI between a given word  $w$  and a class  $c$  can be defined as the level or degree of co-occurrence between the given word  $w$  and the class  $c$ . The mutual information is defined as the proportion of expected co-occurrence of class  $c$  and word  $w$  based on mutual independence and the true co-occurrence which is given in the equation below:

$$M(w) = \log \left( \frac{F(w) \cdot p_i(w)}{F(w) \cdot p_i} \right) = \log \left( \frac{p_i(w)}{p_i} \right) \quad (1)$$

The value indicates how much a particular feature has influence on a class  $c$ . If the value is greater than zero, the word  $w$  is said to have positive correlation with class  $c$  and has negative correlation when the value is less than zero.

### 3.4 Chi-square Method

This method computes a score in order to determine whether a feature and a class are independent. This particular test is a statistic which determines the degree or the extent to which these two are independent. This method makes an initial assumption that the class and the feature are independent and then computes score; with large value indicating they are dependent. Let the number of given documents be denoted as  $n$ ,  $p_i(w)$  indicates the conditional probability of class  $i$  for the given documents which have word  $w$ ,  $P_i$  and  $F(w)$  are the global fraction of documents that has class  $i$  and word  $w$  respectively. The chi-square statistic of the word between  $w$  and class  $i$  is given as

$$M\chi_i^2 = \frac{n \cdot F(w)^2 \cdot (p_i(w) - P_i)^2}{F(w) \cdot (1 - F(w)) \cdot P_i \cdot (1 - P_i)} \quad (2)$$

$\chi_i^2$  is considered to be better than PMI because  $\chi_i^2$  value is normalized and therefore are more comparable across terms in the same type [13].

### 3.5 Multi-class Sentiment Classification

#### 3.5.1 Naive Bayes Classifier

The NB classifier is a simple and commonly used probabilistic classifier for text. This model calculates the posterior probability of a class by making use of word distribution in the given document. This is one of the baseline methods which uses BOW representation that ignores word order. The Bayes theorem is used in order to predict the probability that a given feature belongs to a particular label which is denoted using the equation below:

$$P(\text{label}|\text{features}) = \frac{P(\text{label}) \cdot P(\text{features}|\text{label})}{P(\text{features})} \quad (3)$$

$P(\text{label})$  is the prior probability of the label,  $P(\text{features}|\text{label})$  is the prior probability that a given feature set is being classified as a label.  $P(\text{features})$  is the prior probability that a given feature. Using the naive property which assumes that all features are independent, the above stated equation can be rewritten as below:

$$P(\text{label}|\text{features}) = \frac{P(\text{label}) \cdot P(f_1|\text{label}) \cdot \dots \cdot P(f_n|\text{label})}{P(\text{features})} \quad (4)$$

#### 3.5.2 Maximum Entropy Classifier

The Maxent classifier which is also called as conditional exponential classifier is a probabilistic classifier that employs encoding in order to convert feature sets which are labeled into vector representation. The resultant encoded vector is used for computing weights for every feature which are subsequently combined to identify the most probable label for the given feature set. The model uses  $x$  weights as parameters which combine to form joint features generated using encoding function. The probability of every label is calculated using the equation below:

$$P(\text{fs}|\text{label}) = \frac{\text{dotprod}(\text{weights}, \text{encode}(\text{fs}, \text{label}))}{\text{sum}(\text{dotprod}(\text{weights}, \text{encode}(\text{fs}, l)) \text{ for } l \text{ in labels})} \quad (5)$$



### 3.5.3 Support Vector Machine

SVM is a linear classifier which is best suited for text data because text data is sparse in most cases. However they tend to correlate with each other and are mostly organized into ‘linearly separable categories’. SVM maps the data to an inner product space in a non-linear manner thus building a ‘nonlinear decision surface’ in the real feature space. The hyper plane separates the classes linearly.

### 3.5.4 Ensemble Learners

Ensemble Learners came into existence with the realization that each of the machine learning approaches performs differently on different applications and there is no conformity on which approach is optimal for any given application. This uncertainty has led to the proposal of ensemble learners which exploit the capabilities and functionalities of different learners for determining the polarity of text. Each of the classifiers which are combined are considered to be independent and equally reliable. This assumption leads to biased decisions and this is considered to be a limitation which is to be addressed. The computational complexity associated with big data is very huge and the conventional state of the art approaches focus only on accuracy. Moreover, these approaches performed well on formal texts as well on noisy data. The major highlight of this paper is to evaluate the performance of ensemble learners in short and informal text such as tweets. We have implemented two ‘instance partitioning methods’ like Bagging and Boosting.

### 3.5.5 Bagging

Bagging also known as bootstrap aggregating [8] is one of the traditional ensemble techniques which is known for its simplicity and improved performance. This technique achieves diversity using bootstrapped replicas of training data [8] and each replica is used to train different base learners of same kind. The method of combination of different base learners is called as Majority Vote which can lessen variance when combined with baseline methods. Bagging is often employed in the cases where the dataset size is limited. The samples drawn are mostly of huge size in order to ensure the availability of sufficient instances in each sample. This causes significant overlapping of instances which can be avoided using an ‘unstable base learner’ that can attain varied decision boundaries [11].

### 3.5.6 Boosting

This technique applies weighting to the instances sequentially thereby creating different base learners and the misclassifications of the previous classifier is fed to

the next sequential base classifier with higher weights than the previous round. The main aim of Boosting is to provide a base learner with modified data subsets repeatedly which results in a sequence of base learners. The algorithm begins by initializing all instances with uniform weights and then applying these weighted instances to the base learners in each iteration. The error value is calculated and all the misclassified instances and correctly classified instances are assigned higher and lower weights respectively. The final model will be a linear combination of the base learners in each iteration [9]. Boosting is mainly based on weak classifiers and we have employed AdaBoost, which works by calling a weak classifier several times.

## 4 Experiments

We have implemented this work in R platform. After the pre-processing steps, two feature selection methods viz. PMI and Chi-square were adopted and the most informative features were chosen. The Chi-square method seemed to produce better results than PMI because of the normalization scheme adopted in Chi-square method. Hence we chose Chi-square method as the feature selection strategy in this work and the features generated were fed to the classifiers. The experimental evaluation consists of three phases:

1. The supervised classifiers were trained using most informative features selected using PMI and Chi-square method.
2. Implementations of the state of the art methodologies like baseline methods viz. NB, SVM and Maxent and also ensemble learners like Bagging and Boosting.
3. Performance evaluation using 10 fold cross validation and calculation of precision, recall, f-measure and accuracy. The dataset was divided into ten equal sized subsets from which nine of them constituted training set and remaining one constituted the test set. This process iterated ten times so that every subset became test set once and the average accuracy is computed.

### 4.1 *Baseline and Ensemble Methods*

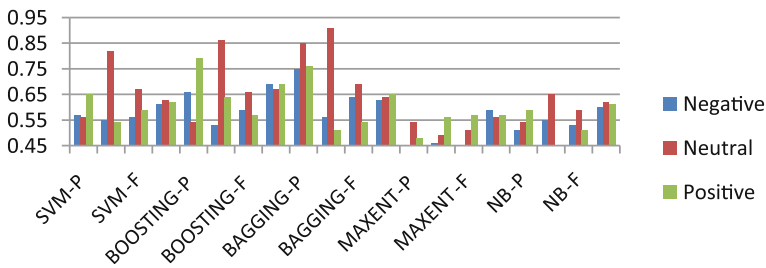
The baseline methods such as NB, SVM and Maxent were investigated and accuracy, precision, recall and f-measure of each of these models were calculated. We have employed 10-fold cross validation for evaluation. The results were compared with ensemble methods such as Bagging and Boosting.

## 4.2 Dataset

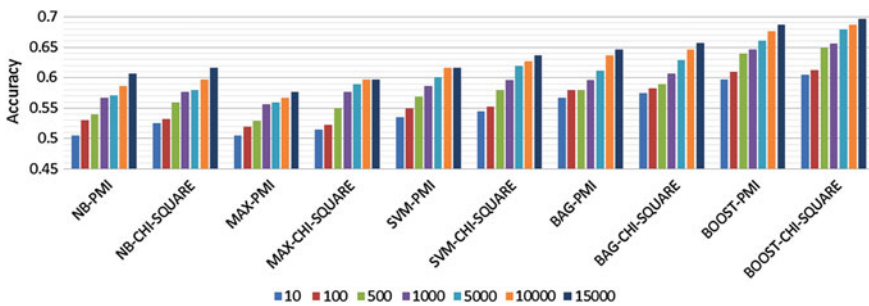
In this work, we have used the twitter dataset provided in SemEval 2013, Task 9 [14]. This dataset consists of tweet ids which are manually annotated with positive, negative and neutral labels. The training set contains 9635 tweets and the testing set contains 3005 tweets.

## 5 Results and Analysis

Figure 2 demonstrates the performance evaluation of baseline methods and ensemble methods using the evaluation metrics like precision, recall, f-measure and accuracy. The graph depicts the values of precision, recall, f-measure and accuracy for each of the positive, negative and neutral classes. The accuracy of sentiment multi class classification that include a neutral class using BOW approaches has not gone beyond 60 % in most cases [15]. The results demonstrate that the ensemble



**Fig. 2** Performance comparison of baseline methods (NB, SVM and MAXENT) and ensemble learners (bagging and boosting) using precision (P), recall (R), f-measure (F) and accuracy (ACC) for each of the three classes: positive, negative and neutral



**Fig. 3** Performance comparison of PMI and Chi-square feature selection methods for different number of features

methods like Bagging and Boosting has achieved significant performance improvement when compared to SVM, NB and Maxent classifiers.

Figure 3 demonstrates the comparison of PMI and Chi-square methods based on accuracy for each of the baseline classifiers and ensemble learners. The results demonstrate that Chi-square method provided improved accuracy than PMI.

## 6 Conclusion

Micro-blogging websites have become very popular and have fuelled significance of sentiment classification. In this work, we evaluated popular and widely used ensemble learners (Bagging and Boosting) for use in tweet sentiment classification into three classes: positive, negative and neutral. The Twitter dataset was used to perform the sentiment classification and the empirical results showed the effectiveness of the above stated Ensemble Learners by comparing with the baseline methods like NB, SVM and Maxent classifiers. The Ensemble Learners tend to produce improved accuracy than the baseline methods. We also incorporated statistical feature selection methods like PMI and chi-square for extracting most informative features instead of using traditional BOW features.

The future directions for this work include: bigger datasets must be used for validating the results obtained in our work because of the imbalanced nature of Twitter datasets. The high computational complexity and running time of the Ensemble Learners need to be tackled by using parallel computing. In addition, the knowledge learnt by the Ensemble techniques are often difficult to interpret by the humans and hence suitable methods must be incorporated in order to improve the interpretability.

## References

1. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 56–65. ACM.2007
2. Whitehead, M., Yaeger, L.: Sentiment Mining Using Ensemble Classification Models: Innovations and Advances in Computer Sciences and Engineering, pp. 509–514. Springer, Netherlands (2010)
3. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. *Ain Shams Eng. J.* **5**(4), 1093–1113 (2014)
4. Liu, B.: Sentiment analysis & opinion mining. *Synth. Lect. Hum. Lang. Technol.* **5**(1), 1–167 (2012)
5. Lek, H.H, Poo, D.C.: Aspect-based Twitter sentiment classification. In: 2013 IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 366–373. IEEE (2013)
6. Fersini, E., Messina, E., Pozzi, F.A.: Sentiment analysis: Bayesian ensemble learning. *Decis. Support Syst.* **68**, 26–38 (2014)

7. Rice, D.R., Zorn, C.: Corpus-based dictionaries for sentiment analysis of specialized vocabularies. In: Proceedings of NDATA (2013)
8. Ortigosa-Hernández, J., Rodríguez, J.D., Alzate, L., Lucania, M., Inza, I., Lozano, J.A.: Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers. *Neurocomputing* **92**, 98–115 (2012)
9. Wang, G., Sun, J., Ma, J., Xu, K., Gu, J.: Sentiment classification: the contribution of ensemble learning. *Decis. Support Syst.* **57**, 77–93 (2014)
10. Breiman, L.: Bagging predictors. *Mach. Learn.* **24**(2), 123–140 (1996)
11. Schapire, R.E.: The strength of weak learnability. *Mach. Learn.* **5**(2), 197–227 (1990)
12. Dietterich, T.G.: Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, vol. 1, p. 15. Springer, Berlin (2000)
13. Aggarwal, C.C., Zhai, C.: *Mining Text Data*. Springer, Berlin (2012)
14. <http://alt.qcri.org/semEval2014/task9/>
15. Wang, H., Can, D., Kazemzadeh, A., Bar, F., Narayanan, S.: A system for real-time twitter sentiment analysis of 2012 US presidential election cycle: In: Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, pp. 115–120 (2012)