International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 9 • Number 45 • 2016

Prediction of Sentiment from Textual Data Using Logistic Regression Based on Stop Word Filteration and Volume of Data

Sukhnandan Kaur[a] and Rajni Mohana[a]

[a]Computer Science and Engineering, Jaypee University of Information Technology Solan, India
E-mail: rajni.mohana@juit.ac.in

Abstract: Sentiment analysis is a process of extracting the opinion of people about any entity from the textual web-data. It is the domain of natural language processing. Nowadays, with the rise in web data and the impact of sentiment analysis in decision support system, the task of sentiment analysis has been gained much attention in the research field. Despite the use of various machine learning algorithms, less concentration is put into the choice of learning algorithms. For the above mentioned context,comparable analysis of various supervised machine learning algorithms is presented in this paper. Extensive performance analysis for supervised machine learning algorithms (Logistic Regression, Naive Bayes and Support Vector Machines) based on the volume of dataset and the impact of stop-word is presented in this paper.This paper mainly focused on the use of logistic regression for binary classification of reviews. The analysis is performed over the benchmark movie review dataset.

Keywords: *Sentiment analysis, Opinion mining, Machine learning,Supervised learning, Naive Bayes, Support Vector Machines, Logistic Regression.*

1. INTRODUCTION

Today business small or large is awash in Data. It is unusual to get the knowledge what all these data is about and the opinion extracted from this huge data on the basis of which authority has to take the decision. Even moderately sized businesses which have dozens of databases, serving many applications encompassing hundreds of gigabytes of data.

In the present web repository, extracting opinion from the textual information available on the web is becoming complex day by day. The reason is the presence of unstructured sentences or informal context available on the social media. As the availability of posting views in a informal manner about any product, person or organization is on the rise, the performance of available machine learning techniques is degrading day by day.It makes the process of opinion extraction more complex. This somehow degrades the performance of decision support system. For getting the reliable decision, researchers have to do a lot of work in the field of sentiment analysis to enhance the accuracy of existing techniques.Many researchers use positive and negative sentiment words to find the opinion of a document or sentences taken randomly from the web. As the research in this field is rising new techniques are enhancing the performance of the system.

Sentiment analysis is an automated technique which tellswhat opinion people have in their mind for a person, product or any place. They express their views through posts on various web pages or review sites in the form of text, audio, etc. With the advent of new techniques, the definition of sentiment analysis is becoming finer.

Thispaper is organised as:section 2 has focused on state-of-art sentiment analysis. In section 3, learning algorithms are described. Section4 contains the description for choice of learning algorithm. In section 5, system design is presented. Section 6 focused on the experiment setup along with the results. Finally in section 7, the whole work is concluded.

2. RELATED WORK

Various linguistic researchers gave automated approaches which are used for determining the polarity at document level, sentence level and aspect level. During late 1990s due to exponential growth of online content, document level sentiment analysis gained attention.Hatzivassiloglou et.al.[1] proposed the technique for sentiment analysis based on the adjective clustering. Barbosa et.al.[2] suggested an approach for sentiment analysis based on Part-of-Speech tagging. Later aspect level sentiment analysis became the need of decision support systems. The problem of finding the entity about which the opinion is being expressedi.e.sentiment classification based on target[3]was catered by Vo et.al.[3]. They have used contextual graph based approach for optimization. The rise in online social data consists of huge number of misspelled words, use of slangs, short text. This makes the online data inconsistent.Xie et.al.[4] proposed a unsupervised technique for extracting sentiments from textual data using linguistic resources.Machine learning algorithms make the task of sentiment classification easy. Most of the automated textual classification approaches are now employed machine learning like Support Vector Machine (SVM) [5], Naive Bayes (NB) [6].Maas et.al. [7] proposed a technique for calculating the semantic likelihood. It combined unsupervised and supervised learning techniques. They proposed vector based model used to deal with semantic and sentiment similarities between different words. It did not work well in the cross domain sentiment analysis. In linguistic world, apart from simple words, there are a number of complex words (like idioms, phrases, proverbs, etc.). These words also contributed to semantic analysis. The consideration of such words is also needed for reliable sentiment analysis. In this paper, we have analysed the results based on various machine learning algorithms. We have also employed logistic regression to observe the performance comparable to various other machine learning algorithms.

3. SUPERVISED MACHINE LEARNING ALGORITHMS FOR SENTIMENT ANALYSIS

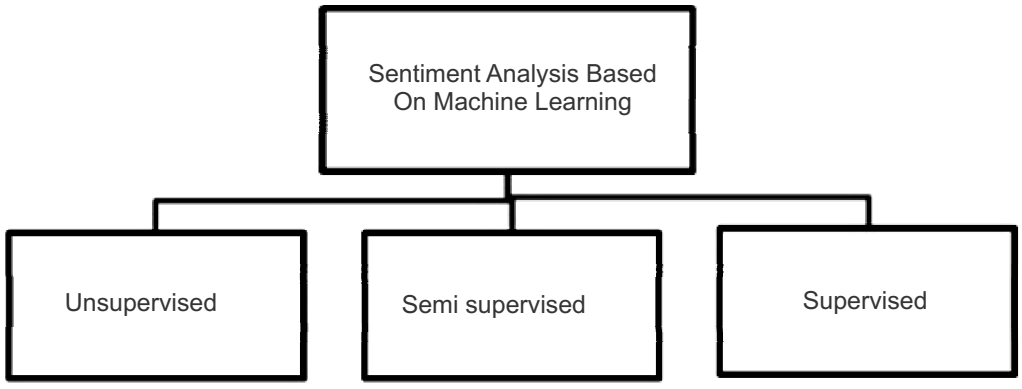

Figure 1: Classification of various machine learning algorithms

In this era of sentiment analysis, machine learning makes thetask of sentiment analysis for taking decision much easier than manual access of data. It is very useful in thefield of sentiment analysis. The need of machine learning is on the hike due to extensive growth in data over the web sphere. This data is very much useful in taking decision regarding any product. Various machine learning techniques are based on training and give the output in the form of prediction. For sentiment analysers, the input is raw data. The output of the analyser is further used by decision support systems. There are numerous learning methods fall into 3 different categories: supervised,unsupervised and semi-supervised.

Training of a system is based on the machine learning model which further used to predict data by developing into on-hand datasets. The classification of state-of-art machine learning methods isshown in Fig.1.

3.1. Supervised learning

In this type of learning, both the features and the class information of training sets are known. This information is used to train a learner for classifying the data sets.Various machine learning algorithms used for supervised learning such as SVM, Naive Bayes, linear regression, logistic regression, etc.Hatzivassiloglou et.al.[1] used supervised learning in their work. Panget.al.[8]used supervised learning for topic classification in sentiment analysis such as Naive Bayes, Max. Entropy and SVM. They found that these machine learning methods are better than human baselines.McCallum et.al.[9] found that for the binary classification of reviews, the performance of Naive Bayes is low as compared to SVM.Feature extraction in these is based on the combination of all active named entities. They reported that these algorithms did not perform as well on sentiment classification as in text classification. Florian et.al.[10]used Hidden Markov Model, A Robust Risk Minimization (RRM) classifier based on regularized winnow methods for named entity extraction in addition to the models used by Pang et.al.[8]. Correlation method for feature extraction used by considering the relationship of bigrams, trigrams and N-grams with the topic by using a distance measure proposed by Liu et.al.[11]. Supervised techniques gave best results in a specific domain oriented environment.

3.2. Unsupervised Learning

On contrary, only the features of training set are known without the class information in unsupervised learning. Therefore, it is difficult to train a learner for classification. It is able to draw some inference from the data. Term contribution, document frequency, term frequency-inverse document frequency count, term variance quality, etc are used for unsupervised learning.Turney et.al.[12] used unsupervised learning for sentiment analysis. For the unannotated data, although we do not know which class an example belongs to, as all of them has the same class space with the labeled data. Rules are built to train the system by using unlabeled data.Florian et.al.[10] proposed agglomerative classifier which was used to do classification based on the active features and their combination. Researchers also have used transformation based learning classifieri.e. Rule based. Chamlertwatet.al.[13]proposed unsupervised machine learning technique based on available lexicon. Unsupervised learning is widely useful, although less accurate as most of the data on the web is unannotated.

3.3. Semi-supervised learning

This is a situation between supervised learning and unsupervised learning. Itstraining set is composed by two kinds of data, one with expected results and another one is without the expected results or un-annotated data. The size of the data which is annotated is not sufficient to be used in supervised learning. Seeding is widely used in semi-supervised learning.Graph based semi supervised learning methods based on minimum cuts proposed by Pang et.al.[14]. Etzioni et.al.[15]proposed bootstrapping method for the minimization of manual labelling of input data. The performance of bootstrapping depends on the seeding done for the extraction at the time of training the system. Riloff et.al.[16]used semisupervised technique, *i.e*. Bootstrapping for annotated data using linguistic clues to extract patterns for subjectivity.

4. SYSTEM DESIGN

The workflow of our system is shown in Fig.2.It consists a corpus which has a collection of positive as well as negative reviews. The system is composed of three main phases:

1. Preprocessing
2. Training and testing
3. Performance Evaluation

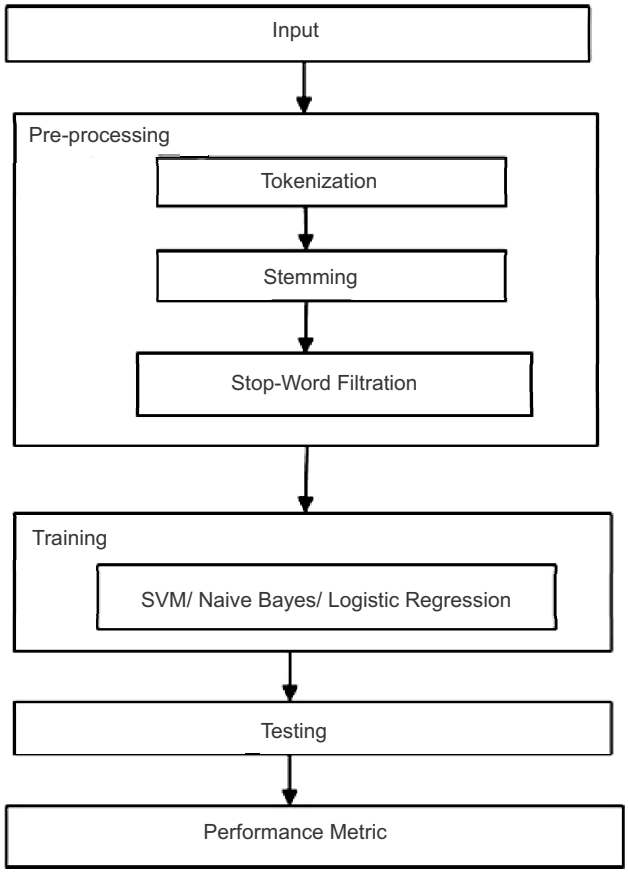

Figure 2: System Design

We experiment with themovie review data set, which is benchmark evaluation data for sentiment analysis. The reviews documents are pre-processed with the Stanford parser[17].

1. **Pre-processing :** Lack of any formalism in social sites, user often use irregular language while posting content over the web. Based on the existing sentiment analysers, the key pre-processing tasks are:
   Tokenization : Tokenization is a kind of pre-processing in a sense; an identification of basicunits to be processed. Stemming— In stemming, root word is formed which reduces the space and time required for processing.
   Stop-Word Removal : The process of removing the auxiliary verbs from the processing is known as stop word filtration.
2. **Training and Testing:** Machine learning methods differ on the information on training sets. For the given dataset, we have separated the whole dataset in two divisions. One is used for training and another one is used for testing. We have used 10 fold cross validation. Machine learning algorithms are used in this paper *i.e*. Naive Bayes, SVM and logistic regression.

3. **Performance Evaluation:** There are the following metrics broadly used to evaluate the system performance. Precision, Recall and Accuracy.

To calculate precision, recall and accuracy the following metrics need to be defined[18]:

1. **True Positives (TP):** Number of positive examples labelled as positive.
2. **False Positives (FP):** Number of negative examples labelled as positive.
3. **True Negatives (TN):** Number of negative examples labelled as negative.
4. **False Negatives (FN):** Number of positive examples labelled as negative

a) **Recall(R):** It is the percentage of named entities present in the corpus that are found by the learning system. It is poor in case of less training data due to which the system is unable to cover all the terms. Recall can be calculated by the equation 1.

$$R = TP/(TP + FN) \quad (1)$$

b) **Precision:** It is the numbers of named entities found by the learning system are accurate. It is found high if it gives correct results. Precision can be calculated by the equation 2.

$$P = TP/(TP + FP) \quad (2)$$

c) **Accuracy(A):** It is defined as the ratio of addition of true positive, true negative and true positive, true negative, false positive, false negative.Accuracy can be calculated by the equation 3.

$$A = (TP + TN)/(TP + TN + FP + FN) \quad (3)$$

5. CHOICE OF MACHINE LEARNING ALGORITHM

Many researchers work in the field of sentiment analysis. There is no hard rule to choose any particular machine learning algorithm. We need to test different algorithms on the same dataset. Based on the precision and recall values of every algorithm, the choice of machine learning algorithm for that particular domain is finalized.In sentiment analysis, most of the work is based on binary classification of dataset i.e. positive or negative. Most of the algorithms used for binary classification are Naive Bayes, Maximum Entropy and Support Vector Machines. The use of Logistic Regression is not tried by the natural language processors for sentiment analysis. In this paper, Logistic Regression is used for binary classification of reviews along with Naive Bayes and SVM. The performance is presented by using Table 1 and Table 2.

Table 1
Experimental Results for movie review dataset 1 using Logistic Regression(LR), Naive Bayes(NB) and Support Vector Machine(SVM)

Vol. 1 DataSet	*Logistic Regression*			*Naive Bayes*			*Support Vector Machine*		
	Recall	*Precision*	*Accuracy*	*Recall*	*Precision*	*Accuracy*	*Recall*	*Precision*	*Accuracy*
With Stop-Word	76.5	76.5	76.5	64	64.02	64.16	77.07	77.14	77.07
Without StopWord	77	77.03	77	64.36	64.38	64.36	77.78	77.81	77.79
Perfomance Difference	0.5	0.53	0.5	0.36	0.36	0.2	0.71	0.67	0.72

Table 2
Experimental Results for movie review dataset2 using Logistic Regression(LR), Naive Bayes(NB) and Support Vector Machines(SVM)

Vol. 1 DataSet	*Logistic Regression*			*Naive Bayes*			*Support Vector Machine*		
	Recall	*Precision*	*Accuracy*	*Recall*	*Precision*	*Accuracy*	*Recall*	*Precision*	*Accuracy*
With Stop-Word	79.95	77.48	79.95	67.05	67.06	67.05	80.15	80.15	80.15
Without StopWord	82.15	82.36	82.15	68.15	68.16	68.15	82.2	82.26	82.2
Perfomance Difference	2.2	4.88	2.2	1.1	1.1	1.1	2.05	2.11	2.05

6. EXPERIMENT RESULTS

Generally from our experiments using machine learning algorithms, we could see that the filtration affects the results to great extent. Therefore, a sentiment analysis without filtration leads to a drop in performance. We did not have a separate dataset for training and testing. Therefore, we used 10 fold cross validation for the analysis of various supervised machine learning algorithms. All the results are based on binary classification of the movie review *i.e.* positive or negative.

Dataset: In our experiment we used two standard datasets. These contain the movie reviews. One dataset (dataset 1) has 1400 movie reviews in which 700 positive and 700 negative reviews. *i.e.*vol.1 dataset. Another dataset(dataset 1)contains the 2000 movie reviews. It has 1000 positive and 1000 negative *i.e.* vol.2 dataset.

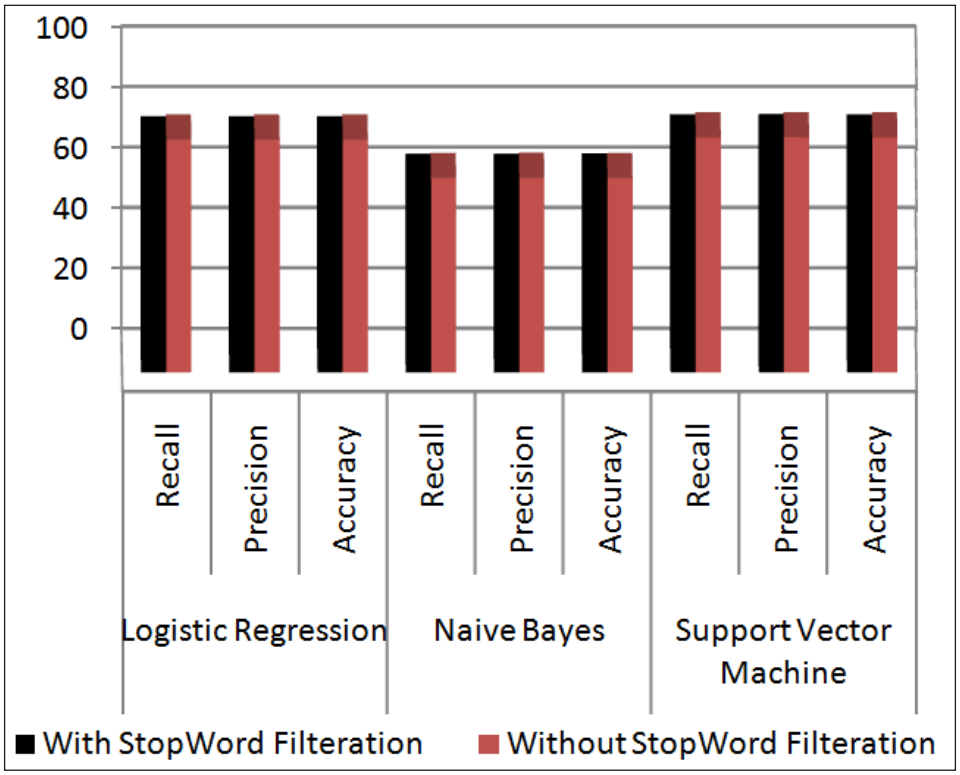

Figure 3: Precision, Recall and Accuracy for dataset vol.1

In table 1, results based on dataset 1 is presented. It shows the Precision, Recall and Accuracy for binary classification of movie reviews. The impact of stop words is shown in Fig. 3 for the same dataset. For our experiment, we have used three machine learning algorithms. Logistic Regression(LR), Naive Bayes(NB) and Support Vector Machines(SVM) are used in the experiment. Table 2 shows the performance of various machine learning algorithms for dataset2. The difference in performance of machine learning algorithms has been calculated based on the numeric value given in last row of the table 1 and table 2. The impact of stop words and volume of training data is shown in Fig.4.

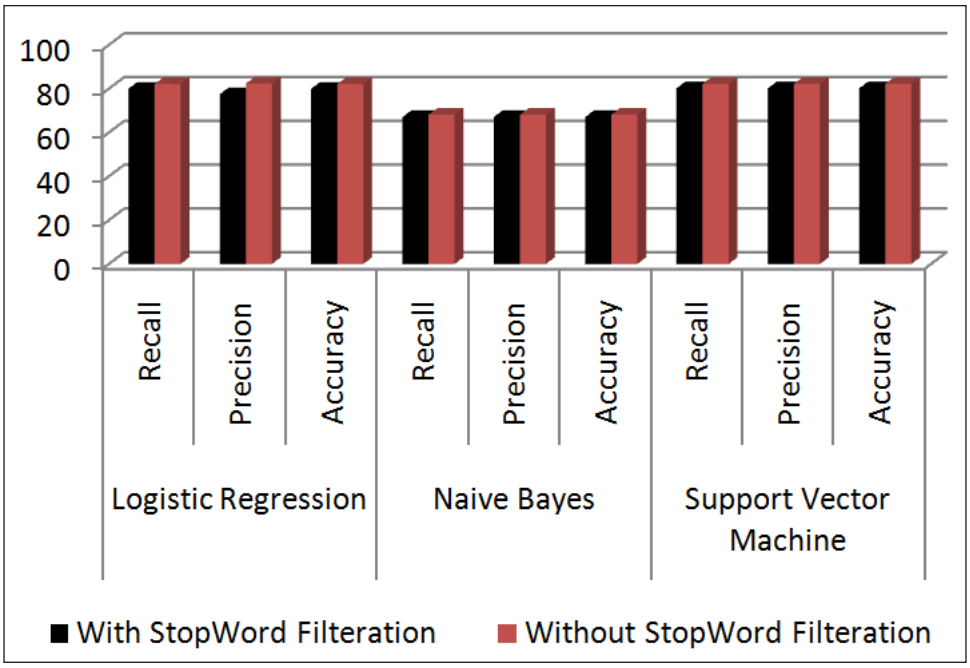

Figure 4: Precision, Recall and Accuracy for dataset vol.2

From table 1 and table 2, it has been seen that the impact of stop-word filtration is more on SVM as its accuracy is increased irrespective to the volume of data. The volume of data also has large impact over the performance metric. The performance of machine learning algorithms has increased with the rise in the training data and by removing stop words from it.

7. CONCLUSION

It is worth trying logistic regression for the movie review dataset. In the literature, researchers have used Naive Bayes, SVM, maximum entropy, etc. After the analysis of various results, we have found that logistic regression outperformed. From various aspects the results have been analysedi.e. the volume of data and stop word filtration. We have presented supervised learning scheme with 10 fold cross validation. In this paper, evaluation of results has showed that Logistic Regressionis better than Naive Bayes for binary classification of the data.

REFERENCES

[1] Hatzivassiloglou V, McKeown KR. Predicting the semantic orientation of adjectives. In: Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics: Association for Computational Linguistics; 1997. p. 174-81.

[2] Barbosa L, Feng J. Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters: Association for Computational Linguistics. p. 36-44.

[3] Vo D-T, Zhang Y. Target-dependent twitter sentiment classification with rich automatic features. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015). p. 1347-53.

[4] Xie S-x, Wang T. Construction of unsupervised sentiment classifier on idioms resources. Journal of Central South University;21:1376-84.

[5] Joachims T. Text categorization with support vector machines: Learning with many relevant features. Springer; 1998.

[6] Srivastava R, Bhatia MPS, Srivastava HK, Sahu CP. Exploiting grammatical dependencies for fine-grained opinion mining. In: Computer and Communication Technology (ICCCT), 2010 International Conference on: IEEE. p. 768-75.

[7] Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C. Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1: Association for Computational Linguistics. p. 142-50.

[8] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10: Association for Computational Linguistics; 2002. p. 79-86.

[9] McCallum A, Nigam K. A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization: Citeseer; 1998. p. 41-8.

[10] Florian R, Ittycheriah A, Jing H, Zhang T. Named entity recognition through classifier combination. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4: Association for Computational Linguistics; 2003. p. 168-71.

[11] Liu L, Kang J, Yu J, Wang Z. A comparative study on unsupervised feature selection methods for text clustering. In: Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on: IEEE; 2005. p. 597-601.

[12] Turney PD. 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: Association for Computational Linguistics; 2002. p. 417-24.

[13] Chamlertwat W, Bhattarakosol P, Rungkasiri T, Haruechaiyasak C. Discovering Consumer Insight from Twitter via Sentiment Analysis. J. UCS;18(8):973-92.

[14] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting on Association for Computational Linguistics: Association for Computational Linguistics; 2004. p. 271.

[15] Etzioni O, Cafarella M, Downey D, Kok S, Popescu A-M, Shaked T, et al. Web-scale information extraction in knowitall:(preliminary results). In: Proceedings of the 13th international conference on World Wide Web: ACM; 2004. p. 100-10.

[16] Riloff E, Wiebe J. Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 conference on Empirical methods in natural language processing: Association for Computational Linguistics; 2003. p. 105-12.

[17] Cer DM, De Marneffe M-C, Jurafsky D, Manning CD. Parsing to Stanford Dependencies: Trade-offs between Speed and Accuracy. In: LREC.

[18] Kaur S, Mohana R. A roadmap of sentiment analysis and its research directions. International Journal of Knowledge and Learning;10(3):296-323.