IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-10, Oct. 2018 E-ISSN: 2347-2693

PREDICTING SENTIMENT FROM MOVIE REVIEWS USING LEXICON BASED MODEL

Binita Verma[1*] , Ramjeevan Singh Thakur[2] , Shailesh Jaloree[3]

Barkatullah University, Bhopal, India[1]
MANIT, Bhopal, India[2]
S.AT.I, Vidisha, India[3]

Available online at: www.ijcseonline.org

Accepted: 12/Oct/2018, Published: 31/Oct/2018

***Abstract*—** Large number of users shares their opinion on social networking sites. So, on the web an enormous quantity of data is generated daily. Usually there is not enough human resource to examine this data. The methods for automatic opinion mining on online data are becoming increasingly. From the past few years, methods have been developed that can successfully analyze the sentiment from digital text. These developments enable research into prediction of sentiment. Sentiment prediction has been used as a tool for movie review prediction. The aim of this work is to explore the use of lexicons to extract the sentiment prediction for a number of movie reviews. In this paper, a comparative analysis of lexicon based models has to predict the sentiments of movie reviews dataset together with evaluation metrics.

***Keywords*—** Movie reviews, Lexicon based model, Predicting sentiment

I. INTRODUCTION

With the rapid advancement of web technology and its growth there is a large amount of data is present on the web for internet users [1][2]. So a huge amount of data generated day by day [3]. Internet has become a platform for sharing opinions online leanings exchanging ideas and so on. There is a large number of social networking sites are available on the web and growing rapidly with their popularity as they allow the people to share their views about a topic have discussed with people or post messages across the world [4]. Sentiment analysis is the field of study that analyze peoples opinion, sentiments evaluation appraisals attitudes and emotions towards entities such as product services organization individual issues events topics and their attributes [5]. Accuracy of the sentiment analysis can be increased by choosing good preprocessing, feature selection and classification technology.

This paper is organized as follows: Section II includes the survey methodology. Section III discuss the various lexicon based approaches of sentiment analysis. In section IV, we describe the implementation of proposed model and finally the conclusion and future work of research is discussed in Section V.

II. RELATED WORK

A lot of work has been done in this area of sentiment analysis from sentiment lexicons.

In some recent work [6][7][8] authors are beginning to attempt to predict the sentiment polarity of reactions of news. Balasubramanyan et al. [7] worked to determining the sentiment polarity of comments in blogs and to predict the polarity of the blog post. They conclude that the community specific PMI method provides a more accurate picture of the sentiment in comments than the general SentiWordNet technique.

Some research in sentiment prediction for news on predicting the movements of stock market [9][10][11]. Sehnal and Song [11] worked on stock market prediction based on sentiments of web users. They develop a model used to make future predictions about stock values. They also able to predict the sentiment with better precision.

Marchand et al. [12] compared seven opinion lexicons on six sentiment datasets. It is found that increasing the lexicon size by semantic expansion as well as assigning an interval value to the words of the opinion lexicon significantly increases the classification performance on short texts.

Nasukawa and Yi [13] introduced the term sentiment analysis in 2003. Instead of classifying the whole document into positive or negative, only used document level sentiment analysis approach to know the polarity of specific subjects of documents.

Megagoda et al. [14] investigated opining mining and sentiment classification studies in three non-English languages to find the classification methods and the efficiency of each algorithm used in these methods. It is found that most of the research conducted for non English has followed the methods used in the English language with limited usage of language specific properties such as morphological variations. The application domains seem to be restricted to particular fields and significantly less research has been conducted in cross domains.

Varghese and Jayasree [15] proposed their work on aspect based sentiment analysis using SVM classifier and SentiWordNet. For the co-reference resolution, Stanford deterministic co-reference resolution system was used.

Ohana and Tierney [16] discussed the problem of automatic sentiment classification of movie reviews with the help of SentiWordNet lexical resource.

Khan et al. [17] proposed domain independent sentiment analysis method using SentiWordNet and a method to classify subjective sentence and objective sentence from reviews and blog comments.

III. LEXICON BASED APPROACH

These approaches calculate emotional orientation of a document from the semantic orientation of words or phrases in the document. These approaches depend on determining number of opinion words annotated with their polarity, strength and semantic orientation.

Lexicon based sentiment analysis of text is a data analysis task performed by employing opinion words and phrases with no prior knowledge opinion words are compiled and collected. Positive and negative words along with opinion phrases are collectively called Opinion Lexicon. Words in the text are evaluated based on opinion lexicon to determine their orientation and henceforth the sentiment of the text. The Lexicon based approach of Sentiment analysis technique is illustrated in figure 1.

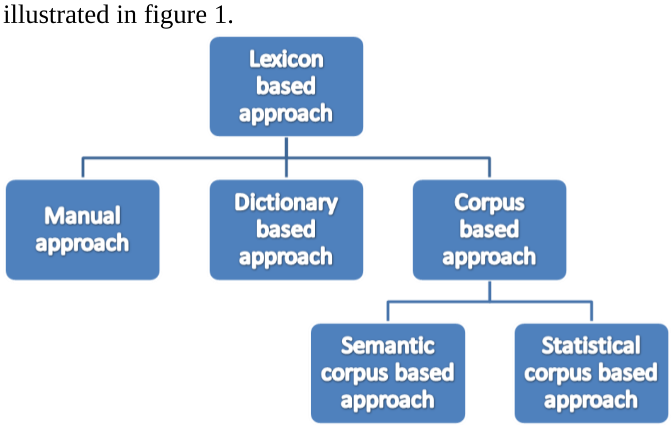

Figure 1. Lexicon Based Approach for Sentiment Analysis

Opinion lexicon generation is a crucial task to lexicon based sentiment analysis process. Generation of opinion lexicon is generally performed using one of three approaches [18][19].

A. Manual Approach

Opinion words are collected manually based on individuals domain knowledge and language understanding. This is a time consuming process. This approach is mostly combined with automated approaches to improve on mistakes done by automated approaches.

B. Dictionary Based Approach

Opinion words with known orientation are collected from lexicographical resources like online dictionary. It uses synonyms, antonyms and hierarchies in opinion lexicons to determine word sentiments. Since there is no knowledge of domain, dictionary based approaches have limitations on identifying context specific sentiment. The dictionary used may be WordNet, SentiWordNet, SecticNet, Sentifull and others.

C. Corpus Based Approach

Corpus based approach exploits the syntactic pattern of co-occurrence words along with opinion words to identify and compile opinion words in large corpus. Corpus based approach eliminates limitation of context specific classification of opinion words in dictionary based approach. However dictionary based techniques are more efficient. Corpus based approach used labeled data. The corpus based approach is performed using semantic approach and statistical approach.

Semantic approach gives sentiment values directly and relies on rules of putting similarity between words. Different kinds of semantic relationship between words are used, computing similar sentiment values for semantically close words.

Statistical approach exploits frequent co-occurring patterns. If word occurs more frequently among positive texts, then its polarity is positive else if it occurs among negative text it's of negative polarity. Two words occurring together frequently in same context have same polarity. This helps find the polarity of unknown words by calculating frequency of co-occurrence with another word.

In lexicon-based model [20], the lexicon is composed of a set of positive and negative opinion words, used to score the opinion sentences positive, negative or neutral [21]. This approach is very popular and requires a scoring function to score every sentence according to the existence of positive and negative words. The lexicon based sentiment analysis model is shown in figure 2. The lexicon based method uses a lexicon a set of positive and negative words combined with a scoring function to determine the sentiment polarity.

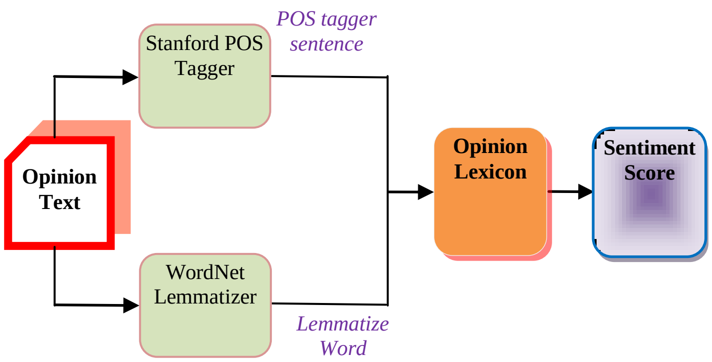

Figure 2. Lexicon based Sentiment Analysis Model

The proposed work is done on movie reviews and evaluation is performed on three English language sentiment lexicons i.e SentiWordNet lexicon, Vader lexicon, and Affin lexicon.

SentiWordNet Lexicon

It is the lexical resource based on WordNet synset or synonym sets and used for sentiment analysis [22]. It assigns three sentiment scores for each synset positive, negative and objective score. The strong word contains higher score and a weak word contains lower score. So sentiment classification is done on scores.

Vader Lexicon

It is based on rule-based sentiment analysis framework and to analyse sentiments in social media[23]. VADER stands for valence aware dictionary and sentiment reasoner. In this lexicon, there are 9000 lexical features from which it was further curated to 7500 lexical features with proper validated scores. Each features rated on a scale from -4 to +4, strong negative to strong positive and 0 for neutral.

Affin Lexicon

It is based on Anew labeled corpus. AFINN – a new wordlist for sentiment analysis on Twitter [24]. A new version Affin-111 contains 2477 words and phrases with their own scores based on sentiment polarity. The polarity depend on positive negative and neutral depend on some numerical score.

IV IMPLEMENTATION OF PROPOSED MODEL

The proposed methodology used to predict the sentiment is shown in figure 3, consist of the following steps: collection of data, text processing and normalization, feature selection, predict sentiment, performance evaluation of model and visualization of the result.

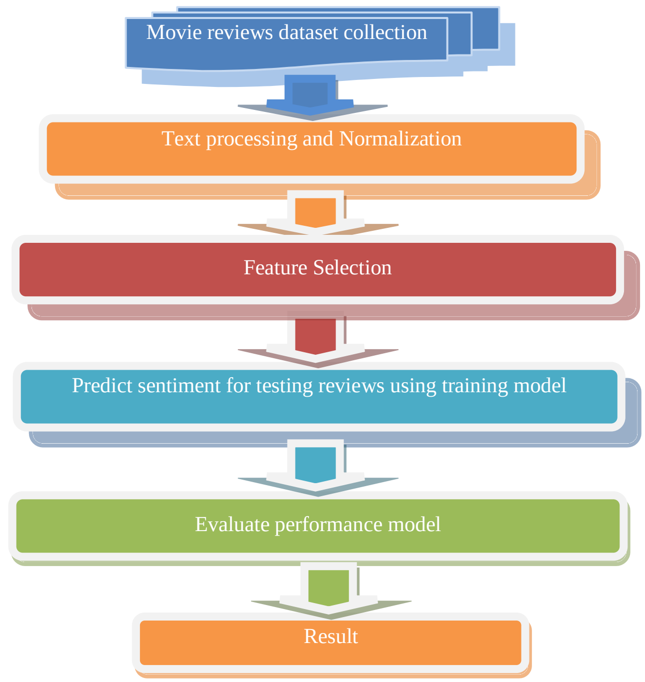

Figure 3. Proposed Model for Sentiment Prediction

A. Data Collection

We use the movie review dataset obtained from internet movie database (IMBD), provided in [25], which is publicly available on Kaggle. For analysis, we obtained 25,000 movie reviews that have been prelabelled with 0 and1 class labels based on reviews, of which 15,000 for training purpose and 10,000 for testing purpose is used. The dataset contains sentiments, id and reviews. We preprocessed the dataset to denote the sentiments by 0 as negative and 1 as positive. We evaluate model performance on the testing data.

B. *Text Preprocessing and Normalization*

The preprocessing of text is necessary for good results in order to reduce noise and provide a structured version of them.

The main components of text preprocessing are:

- Cleaning text: text contains unnecessary contents like HTML tags need to make sure to remove them before extracting features.
- Removing accented characters: when we work with English language we make sure that characters with any other format especially accented characters are converted into ASCII characters.
- Expanding contractions: contractions means shortened version of words. For example do not write as don’t and I would written as I’d.
- Stemming and lemmatization: words that can be created by adding suffix or prefix to the stem word create new words, it is called inflection. The reverse process of

obtain the base form of a word called as stemming. For example walking walked and walker all these words derived from walk.
- Removing Stopwords: words like a, of, and, the, an, and so on do not contribute to the sentiment.

C. Feature based Sentiment Analysis

In lexicon based approach, after the pre-processing of movie reviews, we remove all Feature selection include better performing models, less overfitting, more generalized models, less time for computation and model training and to get a good insight into understanding the importance of various features in your data. In this section, we used Threshold-Based method for feature selection.

In Threshold-Based method, we use some form of cut-off or thresholding for limiting the total number of features during feature selection.

The feature extraction module is responsible for extracting features from the text [26] of movie review can appear as a single word or phrase. Feature based sentiment analysis include feature extraction sentiment prediction sentiment classification and summarization. Feature extraction identifies the product features [27]. Sentiment prediction identifies the word in the sentence containing sentiment or opinion based on sentiment polarity [28]

D. Predict Sentiment for Testing Reviews

Our model work in a movie review tags each word with its corresponding POS tag, extracts sentiment scores for any matched synset token based on its POS tag and finally aggregates the score. Proposed models work on a movie reviews to predict the sentiment of all our test reviews and evaluate its performance.

E. Performance Parameter

We evaluate the sentiment prediction performance of the models on our entire test movie reviews dataset. Movie review mining is more challenging reviews than other dataset review because real life world and ironic terms are mixed in movie reviews. For example unpredictable terms indicate negative opinion but it gives positive opinion for movie reviews. The performance of sentiment analysis is calculated by using confusion matrix Table 1, which is generated when algorithm is implemented on dataset. Various performance measures are used that are accuracy, precision, recall and F1-Score. We have used standard IR performance measures in which we consider TP (True Positive), FP (False Positive),TN (True Negative), and FN (False Negative).

TP= Actual positives that are correctly identified.
TN= Actual positives that are correctly identified.
FP= Incorrect positive predictions.
FN= Incorrect negative predictions.

Table 1. Confusion matrix

		Predicted value	
		Positive	*Negative*
Actual value	*Pos-itive*	**TP** True Positive	**FN** False Negative
	Neg-ative	**FP** False Positive	**TN** True Negative

Accuracy - It is the ratio of total correct prediction over total population, in other words ratio of sum of TP and TN over total positive and total negative instances. Accuracy is calculated by equation (1).

$$\text{Accuracy} = \frac{\text{Total correct prediction}}{\text{Total population}} = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots(1)$$

Precision- it is the ratio of true positive records considering all sentences. The precision is calculated by equation (2).

$$\text{Precision} = \frac{TP}{TP+FP} \quad \ldots\ldots\ldots\ldots(2)$$

Recall - It is the ratio of true positive records considering only positive sentences. It is calculated by equation (3).

$$\text{Recall} = \frac{TP}{TP+FN} \quad \ldots\ldots\ldots\ldots(3)$$

F1 Score - It is the harmonic mean of precision and recall. It is calculated by equation (4).

$$\text{F1-Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{Precision} + \text{recall}} \quad \ldots\ldots\ldots\ldots(4)$$

Tabel 2. Confusion matrix of different lexicon

Confusion matrix for SWN Lexicon			
N=10,000	*Predicted Positive*	*Predicted Negative*	Total
Actual Positive	4624	389	5013
Actual Negative	3695	1292	4987
Total	8319	1681	10000

Table 2. Continued

Confusion matrix for VADER Lexicon			
N=10,000	*Predicted Positive*	*Predicted Negative*	Total
Actual Positive	4210	803	5013

Actual Negative	2297	2690	4987
Total	6507	3493	10000

Confusion matrix for Affin Lexicon			
N=10,000	*Predicted Positive*	*Predicted Negative*	Total
Actual Positive	4215	798	5013
Actual Negative	2153	2834	4987
Total	6368	3632	10000

In Affin lexicon based model, we get accuracy of 70% and F-score of 74%, which is quite better. From the confusion matrix of different lexicon Table 2, we can clearly see that quite a number of negative sentiment based reviews have been misclassified as positive (2153) and this leads to the lower recall of 57% for the negative sentiment class. Performance for positive class is better with regard to recall, where we correct predicted 4215 out of 5013 positive reviews, but precision is 66% because of many wrong positive predictions made in case of negative sentiment reviews. I used the threshold of >=1.0 to determine if the overall sentiment is positive else negative.

In SentiWordNet lexicon based model, we get sentiment prediction accuracy of 59% and F1-score of 69%, which is definitely a step down from Affin based model. We have lesser no. of negative sentiment based reviews being misclassified as positive; the other aspects of the model performance have been affected. A threshold of >=0 has been used for the overall sentiment polarity to be classified as positive and < 0 for negative sentiment.

In Vader lexicon based model, we get accuracy of 69% and F1-Score of 73%, which is quite better than SentiWordNet but less than Affin based model. We can see that, correct predicted 4210 out of 5013 positive reviews. We get Precision 65% and recall 84% for positive sentiment class. Vader recommends using positive sentiment for aggregated polarity >=0.5, neutral between [-0.5, 0.5] and negative for polarity < -0.5. we use threshold of >=0.4 for positive and <0.4 for negative in movie dataset.

F. *Experimental Result and Discussion*

We use python for experimentation. Python is one of the best programming languages when it comes to machine learning and textual analytics. It is easy to learn, open source and effective in catering to machine learning requirements like processing large data sets [29]. The experiments are implemented on 2 GB RAM, Pentium(R) Dual - Core CPU with 3.00 GHz having window 7 operating system and 100 GB hard drive. In this section, we present the experimental details of proposed system. Our proposed model has to predict the sentiments from test dataset movie reviews, and evaluate our performance of proposed model.

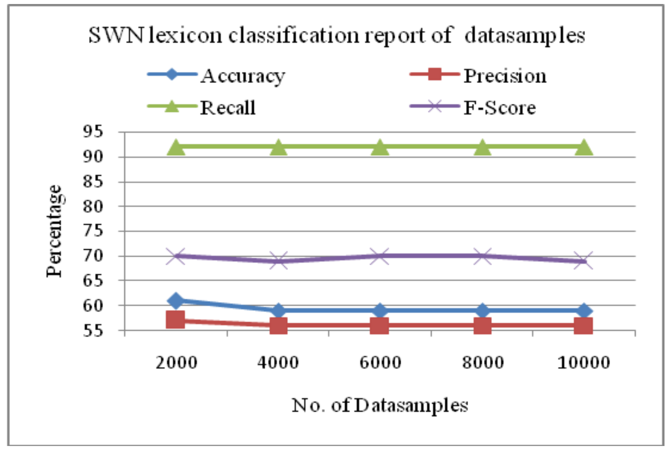

Figure 4. SWN lexicon classification report of different no. of data samples

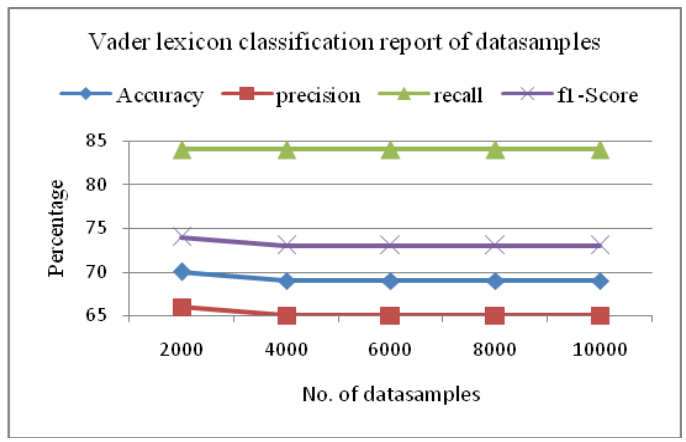

Figure 5. Vader lexicon classification report of different no. of data samples

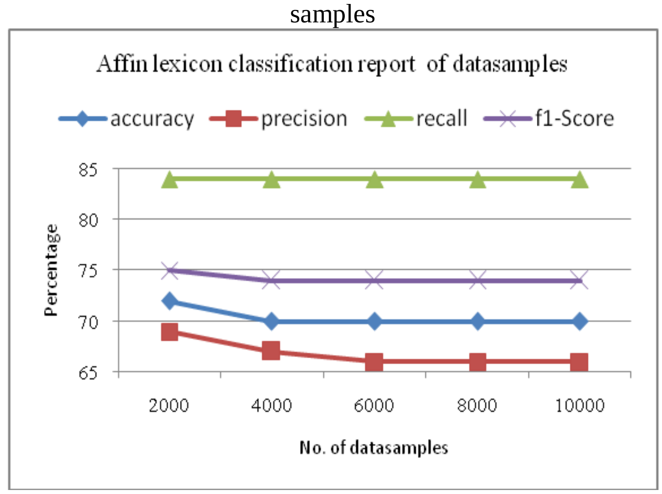

Figure 6. Affin lexicon classification report of different no. of data samples

The accuracy couldn't vary with respect to the number of data samples as shows in the figure 4, figure 5, and figure 6. The ratio of accuracy and amount of data samples doesn't varying as we have taken smaller as 2000 samples and larger has 10,000 samples. So, we can say that if we take much larger samples, the results will be comparable. The performance evaluation metrics of different lexicons is show in table 3.

Table 3. Performance evaluation metrics of different lexicons

LEXICONS	Accuracy %	Precision %	Recall %	F-score %
SWN	59	56	92	69
Vader	69	65	84	73
Affin	70	66	84	74

Affin lexicon consistently obtained the best performance on dataset, among other lexicons. The visualization of the comparison of sentiment analysis model performance is shown in figure 7.

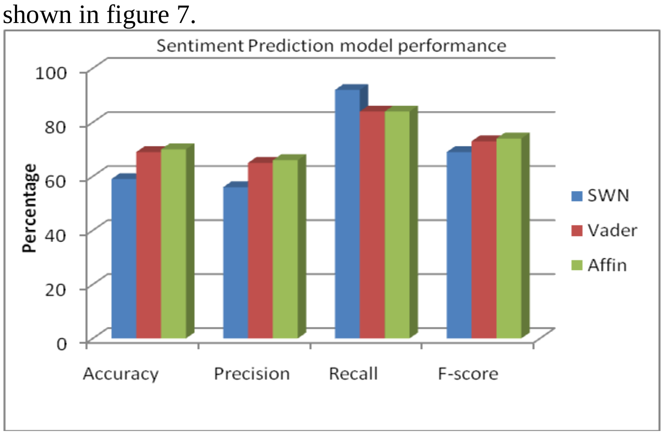

Figure7. comparision of sentiment analysis model performance

V. Conclusion and FutureWork

We evaluate sentiment prediction performance of our proposed models on movie review dataset. The result obtained in term of different evaluation metrics such as accuracy, precision, recall and F1-Score depict the effectiveness of proposed models .From the visualization of results and Table, it is clear that Affin lexicon performs the best among the lexicon based models for movie reviews test data. SWN lexicon model have high recall and low precision that indicates they have a tendency to make more wrong predictions or false positive. In future, we will evaluate more sentiment analysis models as we experiment with supervised machine learning techniques.

References

[1] S. Vishal A. Kharde and S.S Sonawane, "Sentiment analysis of Twitter data: A survey of Techniques", International Journal of Computer Applications, Pp.975-8887, vol. 139 No.11, April 2016.

[2] Pushpendra Kumar and Ramjeevan Singh Thakur, "Recommendation system techniques and related issues: a survey", International Journal of Information Technology, Vol.10 (4), pp. 495–501, 2018.

[3] Vinod Kumar, Pushpendra Kumar and R.S. Thakur, "A brief Investigation on Data Security Tools and Techniques for Big Data", International Journal of Engineering Science Invention, Vol. 6(9), PP. 20-27, 2017.

[4] Pushpendra Kumar and R. S. Thakur, "A Framework for Weblog Data Analysis Using HIVE in Hadoop Framework", In: Proceedings of International Conference on Recent Advancement on Computer and Communication, Lecture Notes in Networks andSystems 34,(2018), https://doi.org/10.1007/978-981-10-8198-9_45

[5] C. Musto, G. Semeraro and M. Polignano, "A Comparison of Lexicon based approaches for Sentiment Analysis of microblog posts", International Workshop on Information Filtering and Retrieval, Pisa, Italy, Dec 2014.

[6] K. Lerman, A. Gilder, M. Dredze and F. Pereira, "Reading the Markets: Forecasting Public Opinion of Political Candidates by News Analysis", In 22nd International Conference on Computational Linguistics, Manchester, UK, pp. 473-480, 2008.

[7] R. Balasubramanyan, W.W.Cohen, D. Pierce and D. P. Redlawsk, "What pushes their buttons? Predicting Comment Polarity from the content of Political blog posts", In Workshop on Language in Social Media, USA, 2011.

[8] R. Balasubramanyan, W. Cohen, D. Pierce and D. Redlawsk, "Modeling Polarizing Topics: When do different Political Communities respond differently to the same news", In 6th International AAAI Conference on Weblogs and Social Media, Dublin, Ireland, 2012.

[9] G. P. C. Fung, J. X. Yu and W. Lam, "News Sensitive Stock Trend prediction", In Advances in Knowledge Discovery and Data Mining : 6th Pacific-Asia Conference, Taipel, Taiwan, pp. 481-493, 2002.

[10] G. Pui Cheong Fung, J. Xu Yu and Wai Lam, "Stock prediction: Integrating text mining approach using real-time news", IEEE International Conference on Computational Intelligence for Financial Engineering, pp. 395-402, 2003.

[11] V. Sehnal and C. Song, "SOPS: Stock Prediction using Web Sentiment", In Seventh IEEE 77 International Conference on Data Mining Workshops, USA, pp. 21-26, 2007.

[12] Y. Marchand, V. Keselj, E. Milios and M. Shepherd, "Quantifying the role of the Opinion Lexicon in Sentiment", Symposium and Workshop on Measuring Influence on Social Media, 2012.

[13] T. Nasukawa and J. Yi, "Sentiment Analysis: Capturing favorability using Natural Language Processing", In Proceedings of the 2nd International Conference on Knowledge capture, Pp. 70–77, ACM, 2003.

[14] N. Medagoda, S. Shanmuganathan and J.Whalley, "A Comparative Analysis Of Opinion Mining And Sentiment Classification In NonEnglish Languages", IEEE International Conference on Advances in ICT for Emerging Regions (ICTer), Pp. 144 – 148, 2013.

[15] R. Varghese and M. Jayashree, "Aspect based sentiment analysis using support vector machine classifier", Advances in Computing, Communications and Informatics (ICACCI)), International Conference on IEEE, Pp. 1581–1586, 2013.

[16] B. Ohana and B. Tierney, "Sentiment Classification of reviews using SentiWordNet", In 9th. IT & T Conference, Pp. 13, 2009.
[17] A. Khan, B. Baharudin and K. Khan, "Sentence based Sentiment Classification from Online customer reviews, FIT , 2010.
[18] B. Baharum, H. L. Lam and K. Khairullah, "A Review of Machine Learning Algorithms for Text-Documents Classification," Journal of Advances in Information Technology (JAIT), Vol. 1, no. 1, pp. 4-20, Feburary 2010.
[19] W. Medhat, A. Hassan and H. Korashy, "Sentiment Analysis Algorithms and Applications: A Survey", Ain Shams Engineering Journal, Vol 5, Issue 4, Pp. 1093-1113, 2014.
[20] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, "Lexicon- based methods for Sentiment Analysis Computational linguistics", Vol 32, Issue 2, PP. 267-307, 2011.
[21] E. Younis, "Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study", International Journal of Computer Applications, Vol. 112, No. 5, pp. 0975-8887, February 2015.
[22] Esuli Andrea and Sebastiani Fabrizio, "SentiWordNet: A publicly available Lexical resource for Opinion Mining", In Proceedings of Language Resources and Evaluation (LREC), 2006.
[23] F. Arup Nielsen, "A New ANEW: Evaluation of a word list for sentiment analysis in microblogs", In Proceedings of the 1st Workshop on Making Sense of Microposts , Pp. 93–98, 2011.
[24] C.J. Hutto and Eric Gilbert, " VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text", Eight International Conference on Weblogs and social Media, 2014.
[25] Andrew Lee Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng and Christopher Potts, "Learning WordVectors for Sentiment Analysis", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol 1, Pp. 142-150, 2011.
[26] Rafael M., D'Addio , Marcos A., Domingues, Marcelo G., and Manzato, "Exploiting feature extraction techniques on users reviews for movies recommendation", Journal of the Brazillian Computer Society, Vol.23, Pp-7, 2017.
[27] Kushal Dave, Steve Lawrence and David M. Pennock , "Mining the Peanut gallery: Opinion extraction and Semantic Classification of product reviews" , In Proceedings of WWW 2003, Pp. 519-528, 2003.
[28] G. Vinodhini and R. Chandrasekaran, "Sentiment Analysis and Opinion Mining : A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 6, june 2012.
[29] I.V. Shravan, "Sentiment Analysis in Python using NLTK", OSFY- Open Source For You, 2016.