A Review of Opinion Mining Techniques

Gaganpreet Kaur, Pratibha Verma and Amandeep Kaur

October 19, 2021

A Review of Opinion Mining Techniques

Abstract

Opinion mining or Sentimental analysis in one of the recent challenge in Natural Language Processing (NLP). Individual express their opinion in different platforms like Facebook, Twitter, Yelp is also a challenging task as the innovation is increasing exponentially. With the growth of social media large amount of data is generated such as comments, review and opinion. But analysis of this data is time consuming and difficult. So there is a need to develop an intelligent system that classify or determine positive, negative, and neutral category. The aim of this paper is to review the concepts and comparison of opinion mining with machine learning, deep learning, transfer learning and Hadoop framework in brief.

Keywords - *Opinion Mining, Machine Learning, Deep Learning, NLP. Transfer learning and Hadoop*

I INTRODUCTION

Sentimental analysis is a contextual mining of text which is used to mining the text and extract the subjective information from the material and helps to business analyst to understand the social sentiment of their business brand or product while monitoring the online data or conversation.. It can be done by lexicon approach (LN) and machine learning approach. Lexicon based approach fails to calculate the data if word not find in the dictionary, but machine learning is easier and efficient but it needs labeled data[1]. According to [2] Sentimental analysis techniques are Sentence level, Document Level, Aspect Level and User Level.

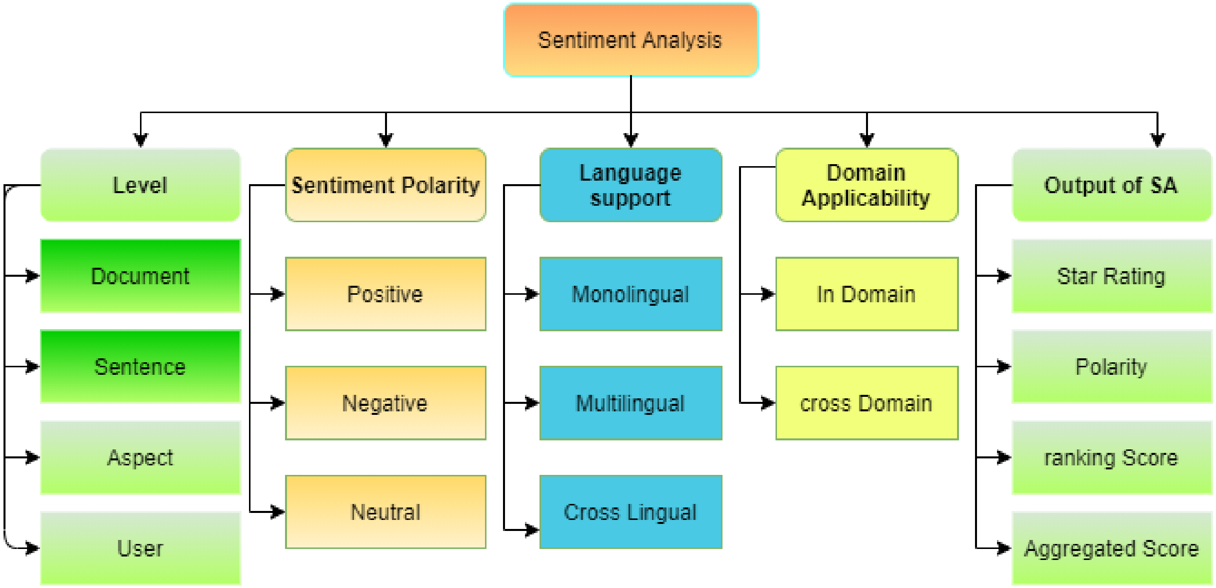


Fig.1Overview of Sentiment Analysis

The sentence level technique classifies the sentiments of each sentence as positive or negative. Main aim to determine the sentiments from a single sentence. The document level technique classifies the sentiments of full document as a single unit. The aspect level technique focuses on properties of an entity. The user level conducts the social interrelation for different parities.

Machine learning and Lexicon based approaches are most common in SA where the machine learning uses testing and training to classify the data. While Lexicon approach are dictionary based which contain predefined positive and negative words [3].While the second approach are used as a dictionary which contain predefined positive and negative words [3]. Fine grained sentiment involves polarity with following categories, very positive, positive, neutral, negative and very negative. These categories then mapped into rating score, for example "very positive" mapped to 5 stars whereas "very negative "mapped to 1 star. For multiple documents, the individual polarities are obtained and then aggregated to give score.

II RELATED WORK

In this section, we will explain the basic analysis procedure and related methods in detail. Figure 2 shows the basic sequential steps.

A Basic Framework of Sentimental Analysis It is a sequential process having various steps .

A. Input:

Data collection is an important part of opinion mining. Data is collected from various sources like Twitter, Facebook, Online Posts, Blogs Micro Blogging and Review sites.

B. Preprocessing

In preprocessing we need to process the collected data:

i) Brackets and number have no meaning in sentimental analysis. They need to remove and are treated as noise.

ii) Tokenization: Tokenization is used to divide the text into smaller components. like "Removal of Extra spaces","Emotions used replaced with their actual meaning like Happy,Sad,"abbreviation like OMG,WTF are replaced by their actual meaning ","Pragmatics handling like happyyyy as happy, gudddd as good and byeeeee as bye etc[4].

(iii) **Stop Word Removal**- Removal of words which are not of any use in analysis like preposition (a,an) and Conjuctions (and, between) [4]etc.

(iv) Stemming- In this process we remove the postfix from each words like "ing","tion" etc[4].used to classify the data or text.

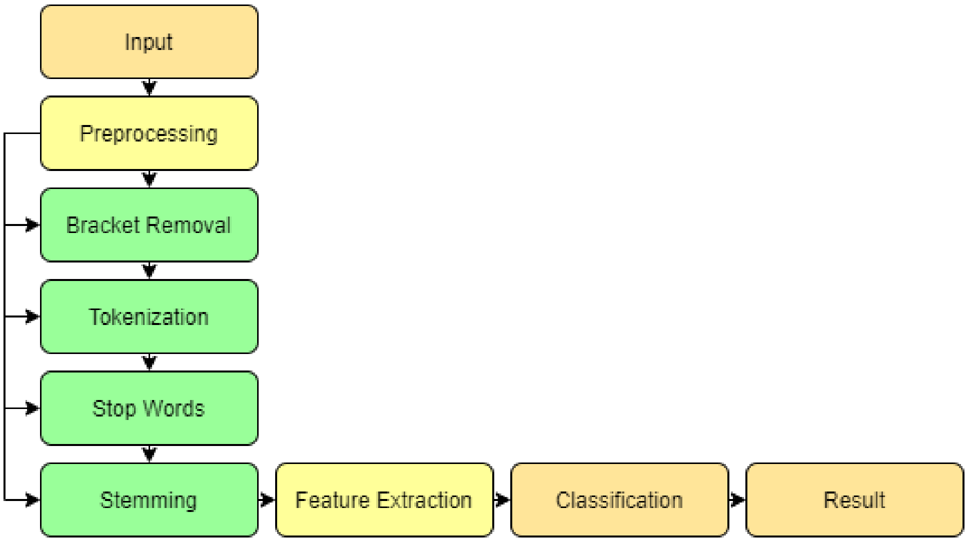


Fig. 2 Basic framework of SA

C. Feature Extraction

Feature is a measurable characteristic of a phenomenon [1]. With feature extraction methods, we extract different features like, adjectives, verb and nouns and later these features identified as positive, negative and neutral polarity of the given data [5].

D. Classification

Several popular and commonly used classification algorithms are used to identify the polarity of user opinion.

B Approaches for Sentimental Analysis

1) **Lexicon Based Approach:** It is a dictionary based approach containing both positive and negative opinion. If the document containing more positive, a positive score is assigned to the required document and document is awarded to negative score if it has high amount of negative words. If the document contains the equal number of positive and negative opinion, a neutral score is given. There are many ways to build and compile a dictionary based method [5].
2) **Dictionary based approach**: Manually small numbers of words (seeds) are gathered with known positive or negative orientation [7]. Searching in WordNet or another dictionary for the synonyms and antonyms. Words found laterly added to the seed list and next iteration begins. The iteration process ends when no more new words can be found.
3) **Corpus Based Approach**: the corpus based approach have dictionaries related to the context or specific domain and it needs large labeled dataset [4].
4) **Machine Learning Based approach**: Machine Learning is concerned with computer programs that automatically improve their performance. The major aim of ML is to allow the systems to learn by themselves through the experience without any kind of human intervention or assistance. The algorithms use computation method to learn directly from data without relying on predetermined equation as a model [9]. The sentiment analysis based on machine learning can be categorized into supervised and unsupervised methods [6].

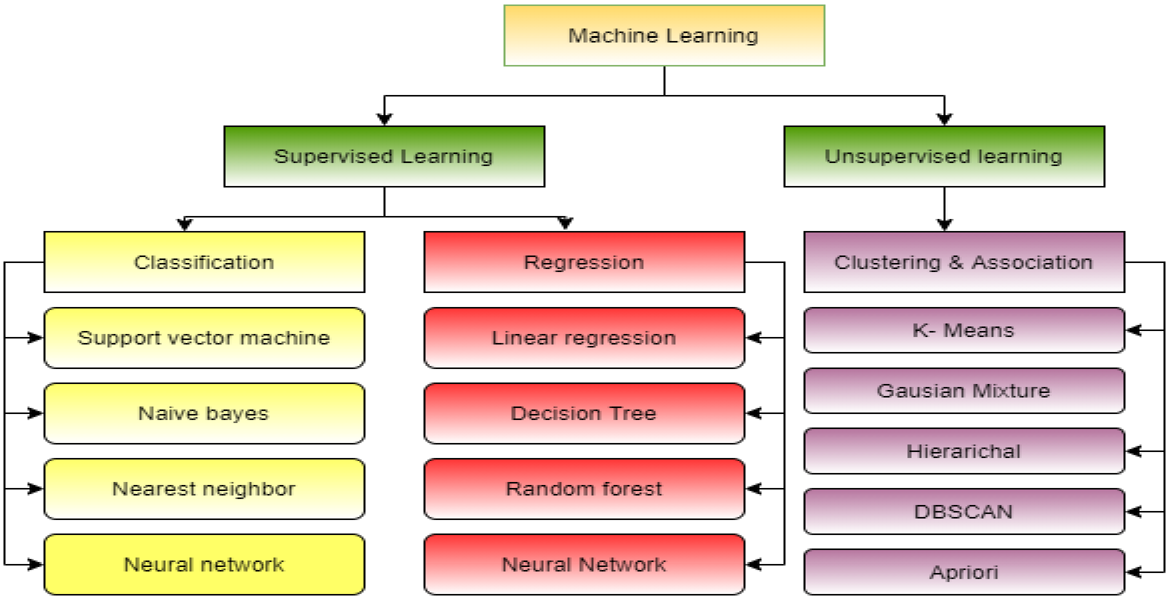
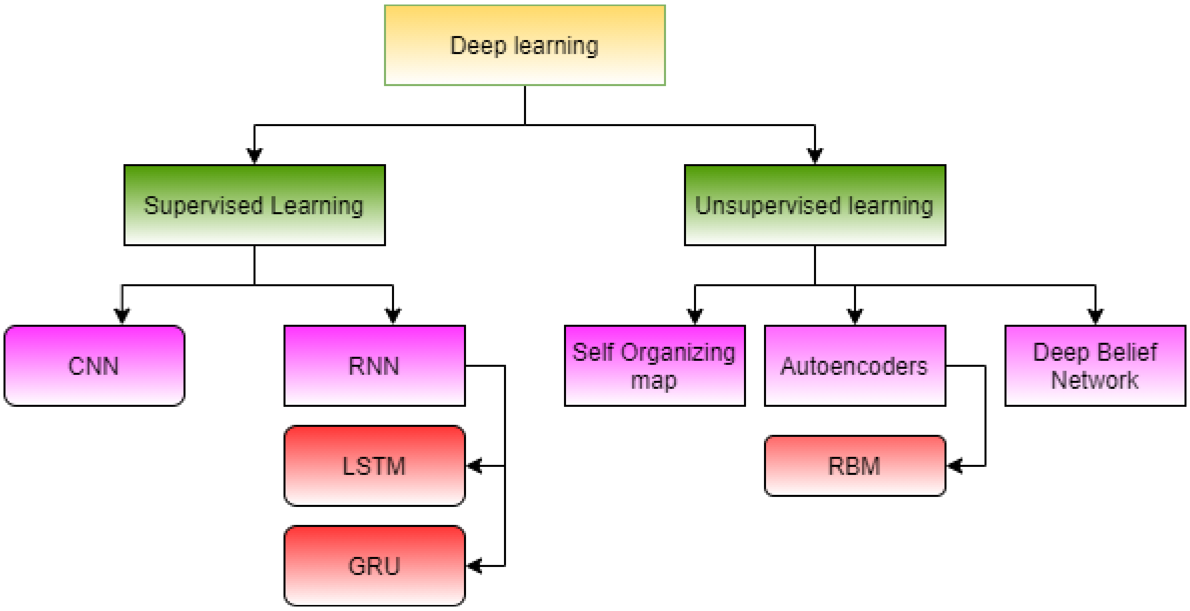


Fig. 3 **Machine Learning Techniques for Sentimental Analysis**

5) **Deep Learning Based approach:** Deep learning is a subtype of Machine learning that uses artificial neural networks to process information as much as a human brain. It is more advanced. When machine learning makes mistakes, human interaction is required however, in deep learning the neural network can learn to correct itself through its advanced algorithm chain. Deep learning models are valuable due to their automatic learning ability. [17]



By using deep learning models sentiment analysis tools utilize the full power. It can be trained to understand text beyond simple definitions, and understand the actual mood and feelings. Sentiment Analysis can be performed by using different deep learning models like CNN (Convolutional Neural Networks), RecNN (Recursive Neural Networks), RNN (Recurrent Neural Networks), DBN (Deep Belief Networks) and HNN (Hybrid Neural Networks).

6) **Transfer Learning Based approach: Transfer** learning (TL) [7] is a machine learning research area that focuses on storing knowledge learned while addressing one problem and applying to a different but related topic. As a result, there is no need to train AI models from scratch. The lack of labeled data makes training NLP models harder. Transfer learning is one of the most effective solutions for dealing with this issue. Transfer learning offers a number of benefits, including reduced training time, improved output accuracy, and the requirement for less training data. For Deep learning, first few of layers are trained to identify features of the problem. So, during transfer learning, you can remove last few layers of the trained network and retrain with fresh layers for target job. Transfer learning is efficient in many computer vision tasks, such as pre-training on ImageNet ,digit recognition ,discovery of cancer subtypes , building utilization and game playing . Transfer learning has also been used in text classification.
The problem of negative transfer is currently one of the most significant barriers to transfer learning. Transfer learning is only effective if the initial and target problems are sufficiently similar for the first round of training to be relevant.
Inductive Transfer Learning: Although the source and target domains are the same, the source and target tasks are

not. The algorithms attempt to use the source domain's inductive biases to help improve the target task. This can be further divided into two subcategories based on whether the source domain contains labeled data or not, multitask learning and self-taught learning.

Transductive Transfer Learning Transfer learning: There are similarities between the source and target tasks in this case, but the contains a large amount of labeled data, whereas the target domain contains none.

Unsupervised Transfer Learning: It is similar to inductive transfer but focuses on unsupervised tasks in the target domain. Although the source and target domains are similar, the tasks are not. This labeled data is not available in either domain.

7) Hadoop Based approach: With more individuals coming online and using e-commerce, the amount of textual information on the internet is growing day by day, making it difficult to analyze massive amount of data efficiently. Hadoop provides a framework that allows collection, storage, retrieval, management, and distributed processing of huge data using cluster of computers and simple programming models. Hadoop framework helps distribute the work among different clustering machines, thus, achieving high performance and each cluster has local storage and can perform local computation. Hadoop is highly performance intensive, scalable, and flexible development framework for parallel processing.

Hadoop addresses all the aspects of Big Data analysis for sentiment determination. Sentiment analysis performance is improved in Hadoop by splitting the data into modules, processing in different machines, reducing response time, and improved fault tolerance by replicating the data. It helps in collection of variety of unstructured data from multiple sources in multiple formats, across domains and efficiently processing them in multi-dimensional fashion. Machine learning algorithms like Naïve Bayes when implemented using MapReduce gives high accuracy for large volumes of data. Machine learning algorithms provided in Mahout scales well for high-dimensional large volume and complex data and can be used in several different applications. Apache Open Source platform using Hadoop also provides reduced cost application to perform sentiment analysis thus help increasing the profit of organization.

II LITERATURE

Paper [9] used semantic orientation calculator (SO-CAL) technique is proposed which work on intensifier and negation. It works on movie review datasets and achieved 76.37 % of accuracy. To detect and classify the sentiments on document level lexicon based approach is used.In [10] uses a three-stage sentiment extraction approach with a document-level approach. Sentiment from datasets is extracted automatically or directly from the internet in the first stage. From this dataset, the second stage will extract positive and negative sets. New document test sets are categorized in the third stage based on the lists gathered in the second stage and the F1 score for positivity is 0.717 and for negative record is 0.622.

In his study [11], the author utilizes sentiment classification to classify Chinese product reviews. Their method was based on unsupervised classification that could educate itself by expanding the vocabulary seed. It began with a single word (good), which was labeled as positive. For sentiment classification, the initial seed was iteratively retrained. The ratio was then calculated using the opinion density criterion.

In [12] a Twitter opinion mining (TOM) framework for tweets sentiment classification is used. According to this hybrid scheme it uses SentiWordNet analysis, emoticon analysis, and an enhanced polarity classifier. The proposed classifier mitigated the sparsity problems by employing various pre-processing and multiple sentiment methods. The experiments were conducted using six datasets demonstrated that the proposed algorithm achieved an average harmonic mean of 83.3%.

In study by [13], the researcher has represented a seven layer framework to analyze the sentiments of sentences. This frame work depends on CNN (Convolution neural network) and Word2vec for SA and to calculate vector representation, respectively. Word2vec have been proposed by Google. The Dropout technology, Normalization and Parametric Rectified Linear Unit (PReLU), have been used to progress the correctness and generalizability of proposed model. The framework was verified on the data set from rottentomatoes.com which contains movie review excerpts' corpus, the dataset consists of five labels positive, somewhat positive, neural, negative and somewhat negative. By comparing the proposed model with previous models such as Matrix-Vector recursive neural network (MV-RNN) and recursive neural network (RNN), the proposed model outperformed the previous models with the 45.5 % accuracy.

Table 1 Analysis of Opinion Mining

Author	Methods	Algorithms	Features	Dataset	Efficiency
E. Kouloumpis, T. Wilson, and J. D. Moore [14] 2011	Supervised Machine Learning	AdaBoost Classifier	POS, Lexicon, Unigram, Bigram and Micro blogging	HASH The Hash tagged and emotion as training dataset	Achieved 0.68 F-measure for HASH and 0.65 by AdaBoost for HASH and EMOT datasets.
Hassan , M. Faizan, Ahsan Hamza, Ahmed , Naeem[15]	Machine Learning	SVM ,Random Forest, Logistic Regression (LR),K-Nearest Neighbor(KNN),	F –Score, Accuracy Score, N gaming, Tokenization, lemmatization and stop word removal	Scientific Citations	The F Score has improved by 9%. The F score is 87%, while 78% from the previous results.
Siva Kumar Pathuri, Dr.N.Anbazhagan, Dr.G.Balaji Prak[16]	Machine Learning	Hybrid BagBooster approach, Naïve Bayes , logistic Regression and SVM	F1 score, Precision , Recall, Unigram, Bigram, Trigram with different size	Review of Mobiles collected from Amazon	When compared to other algorithms, this paper obtained an accuracy of 94 % using the Hybrid Bag-booster algorithm.
Rubeena Parveen,Neelesh,Pradeep Tripathi[17]	Machine Learning Ensemble	Naïve Bayes, SVM KLearn, Multinomial NB, GaussianNB, Bernoulli NB	POS,Tokenization, Tagging,	Movies review from corpora community.	This paper got 91% accuracy with proposed algorithm in which they merge weak leto get better result.

Qiongxia Huang, [18](2017)	Deep Learning	CNN -LSTM	F1 Score	sina micro-blog provided by the seventh COAE (Chinese Orientation Analysis and Evaluation)	The model, which consists of 1 layer CNN and 2 layers LSTMs stacked in order, achieved an accuracy of 87.2 % which is 4% higher than the LSTM and F1 score is 0.868, which is 4% higher than CNN-LSTM.
Mehmet Umut, Ilhan Aydin Salur[19](2020)	Hybrid Deep Learning	CNN, LSTM, BiLSTM, and GRU	.URL distribution, Fast Text , F1 , kappa , Recall	Turkish Twitter messages about GSM operaters	M-7-B achieves the highest accuracy using the word embedding approach, while M-1-A achieves the highest accuracy using the character-level embedding approach. The accuracy of the M-1-A, M-7-B, and M-Hybrid models is 75.73%, 80.03%, and 81.77%, respectively. The performance of the M - hybrid model is higher than that of the other basic models.
Hossein Sadr , Mir Mohsen Pedram , Mohammad Teshnehlab[20] (2020)	Heterogeneous deep neural networks Multiview classifier	CNN-RecNN + KCCA CNN-RecNN + SVM2K CNN-RecNN + MVMED CNN-RecNN + SMVMED	Intermediate representation	Stanford Sentiment Treebank as SST1, SST2	It shows multi-view classifier increase the accuracy than single-view classifiers. CNN-RecNN +SMVMED outperforms 91.93%
Ngoc C. Leˆ, Nguyen The La, Son Hong	Transfer leaning	BERT	Precision Recall F1score	Vietnamese dataset of Hotel and Restaurant	This method improves the F1 score by 10%, as well as precision and

Nguyen, Duc Thanh Nguyen[21](2020)					recall, which improves overall accuracy over previous models.
Salih Emre Akın Tugba Yıldız[22](2019)	Transfer leaning fine-tuning technique and deep learning architecture	Regular Long Short-Term Memory (LSTM) with dropout.	unigrams, bigrams , (POS) tags,	Turkish articles from Wikipedia.	According to this paper, with-TL model outperformed without-TL. While the accuracy of the restaurant dataset is 90.1 % when using TL and 86.8 % when not using TL, using TL yielded 2.3% better results.
E. Omara, M. Mosa and N. Ismail.[28](2019)	Deep and Transfer learning	Deep CNN model	Character level representation	Emotion Tone Data set and Arabic tweets data set	Deep CNN outperforms other models (nearly 1.3 percent) in sentiment analysis. When compared to traditional machine learning models, emotion detection with deep CNN and transfer learning achieves an accuracy of 95.24 %.
K.Sridharan, G.Komarasamy, S.Daniel Madan Raja[31](2020)	Hadoop framework	Random Forest	Accuracy, Precision, Recall,F measure	Clothing, shoes, and jewelry dataset from Amazon .	The accuracy of RF increases as the number of trees increases. The accuracy improves by 5.77%, 3.14%, 1.85%, and 0.58% when the number of trees used is 250, compared to 50, 100, 150, and 200, respectively.
Rathor, S.[32] (2020).	Apache framework	Pig and Flume tool	Positive , Negative , Neutral	Twitter large data set	It provides efficient result with lower computation time.

RESULTS/ Conclusion

Traditional approaches, such as lexicon-based approaches, take a long time to complete. They also have a hard time generalizing to other domains or sectors. Even with standard machine learning algorithms, the most time-consuming processes are feature engineering and feature extraction. As a result, deep learning alleviates the burden of feature creation because as the network learns, it produces the features on its own. The lack of labelled data makes it difficult to train NLP models. Transfer learning is one of the most effective solutions to this problem.There is no need to start from scratch when training AI models. Transfer learning has several advantages, including shorter training times, higher output accuracy, and the need for less Sentiment analysis can be carried out without sacrificing accuracy or speed. It can scale to larger data sets while maintaining performance. Hadoop-implemented machine learning algorithms are simpler and modular, with few lines of code. [1] Hadoop-implemented sentiment analysis is less complex, more easily extendable, and provides high performance at a lower cost. However, advanced processing techniques are required for such operations due to the 5 V model of big data characteristics: volume, variety, velocity, variability, and veracity. training data. Hadoop is used because it is a free and open-source software that is effective for storing and manipulating large amounts of data in distributed patterns.

REFERENCES

[1] Atiqur Rahman, Md. Sharif Hossen,” Sentiment Analysis on Movie Review Data UsingMachine Learning Approach “International Conference on Bangla Speech and Language Processing (ICBSLP), 27-28 September, 2019.978-1-7281-5242-4/19 ©2019 IEEE.

[2] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li, “User-level sentiment analysis incorporating social network,” In Proc. Of ICKDDM, IEEE, pp. 1397-1405, 2011.

[3] X. Fan, X. Li, F. Du, Xin Li, Mian Wei, “Apply word vectors for sentiment analysis of APP reviews,” In Proc. of ICSI, IEEE, 2016.

[4] Prerna Mishra, Dr. Ranjana Rajnish, Dr.Pankaj Kumar,” Sentiment Analysis of Twitter Data:Case Study on

[5] Digital India,”In 2016 International Conference on Information Technology(InCITe)The Next Generation IT –Summit.

[6] Mitali Desai, A. Mehta,” Techniques for Sentiment Analysis of Twitter Data: A Comprehensive Survey”, International Conference on Computing, Communication and Automation (ICCCA2016), ISBN: 978-1-5090-1666-2/16/$31.00 ©2016 IEEE.

[7] Reshma Bhonde , Binita Bhagwat , Sayali Ingulkar , Apeksha Pande,” Sentiment Analysis Based on Dictionary Approach”, International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015, PP 51-55 ISSN 2349-4395 (Print) & ISSN 2349-4409

[8] Siva Kumar Pathuri,Dr.N.Anbazhagan,Dr.G.Balaji Prakash,” Feature Based Sentimental Analysis for Prediction of Mobile Reviews Using Hybrid Bag-Boost algorithm, IEEE 7th International Conference on Smart Structures and Systems ICSSS 2020..

[9] A. Harb, M. Plantié, G. Dray, M. Roche, F. Trousset, and P. Poncelet, "Web Opinion Mining: How to extract opinions from blogs?," in Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, 2008: ACM, pp. 211-217.

[10]T. Zagibalov and J. Carroll, "Unsupervised classification of sentiment and objectivity in Chinese text," in Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I, 2008.