International Interdisciplinary Virtual Conference on
'Recent Advancements in Computer Science, Management and Information Technology'
International Journal of Scientific Research in Computer Science,
Engineering and Information Technology | ISSN : 2456-3307 (www.ijsrcseit.com)

Comparative Study of Lexicon Based and Machine Learning Based Approaches for Sentiment Analysis

Asst. Prof. R. B. Ghayalkar[1]*, Dr. D. N. Besekar[2]
*[1]Assistant Professor, [2]Professor (Retired)
Department of Computer Science, Shri R. L. T. College of Science, Akola, Maharashtra, India

ABSTRACT

Sentimental Analysis is the task of Natural Language Processing (NLP). Sentiment analysis is an emerging technology that aims to explore people's review, feedback, opinion toward any entity in different fields, such as product review analysis, public opinion analysis for decisions making. In this paper comparatively studied the different approaches for Sentiment Analysis and its performance to determine the sentiments of user's data.

Keywords: NLP, Sentiment Analysis, Lexicon Based, Machine Learning

I. INTRODUCTION

Sentiment analysis or opinion mining sometime use interchangeably, It is the field of NLP that analyses people's opinions, sentiments, review, feedbacks, attitudes or emotions towards any entities such as products, services, organizations, individuals, events, issues, any topics etc. for making right decision.

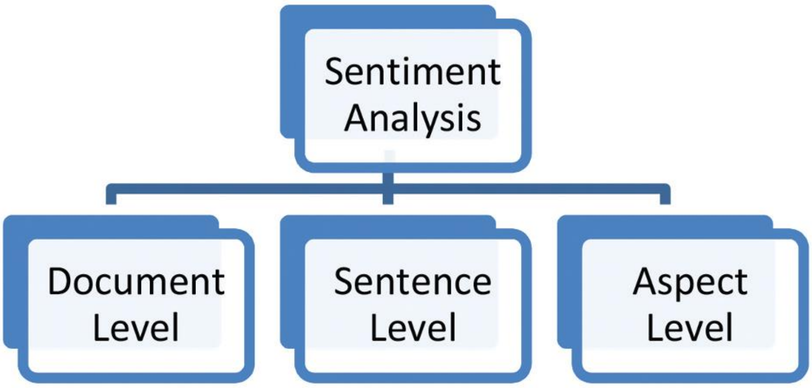
Sentiment analysis is a NLP problem. It touches every aspect of NLP, e.g. reference resolution, negation handling, and word sense disambiguation, which add more difficulties to solved problems in NLP. Sentiment analysis offers a great platform for NLP researchers to make tangible progresses on all fronts ofNLP with the potential of making a huge practical impact

There are three levels classification of sentiment analysis

Document level: The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment.

Sentence level: The task at this level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion.

Entity and Aspect level: Both the document-level and sentence-level analyses do not discover what exactly people liked and did not like. Aspect level performs finer-grained analysis. Aspect level was earlier called *feature level* (*feature-based opinion mining and summarization*). It is based on the idea that an opinion consists of a *sentiment* (positive or negative) and a *target* (of opinion). [1]



There are three approaches for analysing the Sentiments: lexicon-based, machine-learning-based, hybrid approaches.

Machine Learning Approach

In artificial intelligence, machine learning is one of its subsections which are proceeding with algorithm that let systems to understand. In machine learning technique it uses supervised learning and unsupervised learning.

Lexicon Based Approach

In lexicon based method it supports a lexicon to achieve sentiment classification through weighting and counting sentiment associated words has to be calculated and labeled. To assemble the viewpoint list there are three major methods are considered: dictionary-based method, corpus-based method and the manual opinion approach.

Dictionary-based classification

This type of classifications the data are collected from manually and the information is searching for synonymsand antonyms of sentiment dictionary. These dictionaries are WordNet dictionary and sentiwordNet dictionary.

Corpus-based classification

This approaches the objectives of dictionaries related to the specific domain. The words are related to statistical and semantic methods that are Latent Semantic Analysis (LSA) and method based on semantics.[2]

II. LITERATURE REVIEW

There has been done lot of work in the last few years, many articles, books and research papers have been written on sentimental analysis

Chuanming Yu et.al, [3], presented in the paper, during experiment four data sets were used to test the SVM model. Authors have compared Maximum Entropy classifier method for feature extraction with SVM method and they have concluded thatSVM Method superior in terms of recall and precision rates.

Raisa Varghese et.al, [4]authors proposed in this paper different approach which bunch up the benefits of Senti-WordNet, dependency parsing, and co reference resolutions are well organized for the purpose of sentimental analysis. This was done by using Support Vector Machine classifier.

AmitGupte et.al, [5] presented in this paper the comparison between most likely used approaches for sentiments like Naive Bayes, Max Entropy, Boosted trees and Random Forest algorithms.

GautamiTripathi et.al, [6] Inthis paperauthors applied SVM and Naïve Bayes classifier in analyzing the movie sentiments. By this categorization they conclude that linear Support Vector Machine outperforms the Naïve Bayesian in case of accuracy.

Deepu S. Nair et.al, [7] demonstrated in the paper how machine learning technique was used to understand the Malayalam movie comment. For classifying the sentiments two machine learning approaches are used; they are SupportVector Machine and CRF along with rule based approach.

Suchita V Wawre1 et.al, [8] compared two most frequently used supervised machine learning approaches SVM and Naive Bayes for sentiment classification of reviews. The result shows that SVM has misclassified more number of data points as compared to Naive Bayes and Naive Bayes approach outperformed the SVM when there are less number of reviews.

Vishal A. Kharde et.al, [9] used lexicon-based methods for classification but it requires small effort in individual labeled text document.

Neha S. Joshi et.al, [10] In the papershown the outline of recommended methods along with its most recent advancements in the same field. As a result, authors concluded that unsupervised machine learning techniques fails to provide better achievement in sentiment classification than that of supervised learning.

AlessiaD'Andrea et.al, [11] described in this paper various tools used in sentimental analysis and some approaches for text classification. In this method they use hybrid approach which uses the aggregation of both lexicon based and machine learning techniques

Shun Yoshida et.al. [12] inthe paperproposed Naïve Bayesian classifier to analyze the sentences. Their experimental result shows that Naive Bayesian classifier model which has acceptable achievement for distinct Social Network Site and for large data set in which it consists of long comments.

Li Yang, et. Al. [13] proposed a new sentiment analysis model-SLCABG, which is based on the sentiment lexicon and combines Convolutional Neural Network (CNN) and attention-based Bidirectional Gated Recurrent Unit (BiGRU).

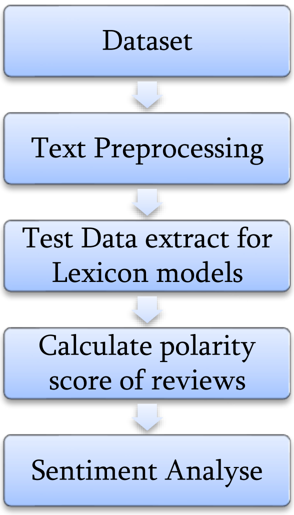
Mahesh B. Shelke et.al, [14] presented in this paper gives a comparative analysis of sentiment analysis performed in various Indian languages, which includes classification techniques which are based on Lexicon, Dictionary, and Machine Learning.

Pei Ke et.al, [15] proposed a context aware sentiment attention mechanism to acquire the sentiment polarity of each word with its part-of-speech tag by querying SentiWord-Net.

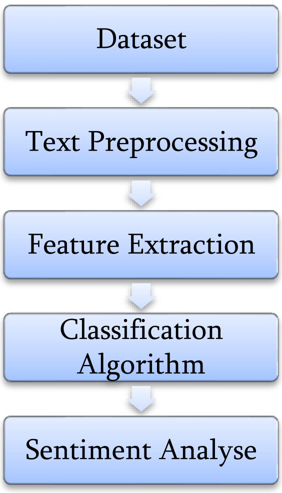
KomangWahyuTrisna et.al, [16] described in this paper, aspect-based sentiment analysis (ABSA). First, task of ABSA, there are three subtasks that we describe namely, aspect term extraction, aspect term categorization, and aspect term sentiment analysis and provided several models for each task from ABSA. Second, describe the deep learning methods that are used to solve the ABSA tasks and then described two popular datasets that used in the ABSA task.

Yuxing Qi et.al, [17] in this paper presented, extracts data regarding Covid-19 from people in the main cities of England on Twitter and separates it into three different stages. First, perform data cleaning and use unsupervised lexicon-based approaches to classify the sentiment orientations of the tweets at each stage. Then, apply the supervised machine learning approaches using a sample of annotated data to train the Random Forest classifier, Multinomial Naïve Bayes classifier, and SVC, respectively.

III. METHODOLOGY


Dataset
Text Preprocessing
Test Data extract for Lexicon models
Calculate polarity score of reviews
Sentiment Analyse

Lexicon Based Approach



Machine Learning Based Approach

IV. PRE-PROCESSING TECHNIQUES

Tokenization: This step breaks the large paragraphs called chunks of text is broken into tokens which are actually sentences.

Normalization: It includes the conversion of all text to either upper or lower case, eliminating punctuations and conversion of numbers to their equivalent words.

Stemming: The stemming process is used to change different tenses of words to its base form this process is thus helpful to remove unwanted computation of words.

Lemmatization: Lemmatization is the process of merging two or more words into single word

Removing Stop Words: Stop words refer to most common words in the English language which doesn’t have any contribution towards sentiment analysis.

Noise removal: The datasets taken comes in raw form. [2]

V. FEATURE EXTRACTION

TF-IDF:

The term frequency-inverse document frequency (also called TF-IDF), is a well-recognized method to evaluate the importance of a word in a document. Term Frequency of a particular term (t) is calculated as number of times a term occurs in a document to the total number of words in the document. IDF (Inverse Document Frequency) is used to calculate the importance of a term.

N-Gram:

N-Gram will form the features of text for supervised machine learning algorithms. These are sequence of n tokens from the given text. Value of n can be 1, 2, 3, and so on. If we consider the value of n to be 1 it is called unigram, for n=2, bigram and for n=3 trigram and so on.

Classification Algorithms

Logistic Regression: This is a popular classification algorithm which belongs to class of Generalized Linear Models. The probabilities describing outcome of a trial is modeled using logistic regression [18]. This algorithm is also called Maximum Entropy.

Naive Bayes: This is powerful algorithm for classification used for classifying data on basis of probabilities. It simply works on Bayes theorem and uses various probabilities to classify data. In Naïve Bayes class with maximum probability is considered to be as the predicted class. It is a fast and highly scalable algorithm. It can also be used on small datasets and thus also gives good results [19].

Support Vector Machine

This is an efficient algorithm for regression as well classification purpose. It draws a hyperplane to separate classes. This algorithm works extremely well with regression, the effect of SVM increases as we increase dimensional space. SVM also perform well when the dimension number is larger than the sample number. [20]

Decision Tree

This algorithm can be used for both regression and classification. The core idea is to divide the dataset into smaller subsets and at the same time tree associated is incrementally created. [21]

K-Nearest Neighbour (KNN)

This algorithm is simple and has applications mainly in pattern recognition; intrusion detection and many more are also there.

Random forest: Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems. [22]

Machine Learning Algorithm	Advantages	Drawbacks
KNN	It is simple and also used for multiclass	It requires more time to categorize when

	categorization of document.	huge number data are inclined. Takes lot of memory for running a process
Decision Tree	This is very fast in learning data set. Easy for understanding purpose	It has problem that it is difficult handle data with noisy data Over fitting of data
Naïve Bayesian	Simple and work well with textual as well as numerical data. Easy to implement Computationally cheap	Performs very poorly when feature set is highly correlated. It gives relatively low classification performance for large data set. Independent assumption of attribute may lead to inaccurate result.
Support Vector Machine	High accuracy even with large data set Works well with many number of dimensions No over fitting	Problems in representing document into numerical vector
Naive Bayes	**It is simple and easy to implement**. It doesn't require as much training data. It handles both continuous and discrete data. It is highly scalable with the number of predictors and data points.	Naive Bayes assumes that all predictors (or features) are independent, rarely happening. This algorithm faces the 'zero-frequency problem
Logistic Regression	Logistic regression is easier to implement, interpret, and very efficient to train. It is very fast at classifying unknown records.	If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.

VI. CONCLUSION

In this paper studied different approaches as well as different techniques for classification have been studies to find out sentiment analysis. Comparatively each approach is suitable according to what type of data that analyse both lexicon based and machine learning based work well but as far this study concern hybrid approach will more comprehensively work.

Future work should be to overcome the challenges of sentiment analysis are work ambiguity, sarcasm, use of Emoji, multimodal data classification, multilingual feedbacks.

VII. REFERENCES

[1]. Bo Pang, Lillian Lee, “Opinion Mining and Sentiment Analysis”, Foundations and TrendsR_ in Information Retrieval, Vol. 2, Nos. 1–2 (2008) 1–135, _c 2008, DOI: 10.1561/1500000001.

[2]. RavinderAhuja, Aakarsha Chug, ShrutiKohli, Shaurya Gupta, PratyshAhuja, “The Impact of Features Extraction on the Sentiment Analysis, International Conference on Pervasive Computing Advances and Applications – PerCAA 2019, ScienceDirect, Elsevier

[3]. Chuanming Yu, “Mining Product Features from Free-Text Customer Reviews: An SVM-based Approach”, iCISE 2009 December 26-28, 2009, Nanjing, China.

[4]. Raisa Varghese, Jayasree M, “Aspect Based Sentiment Analysis using Support Vector Machine Classifier”, 978-1-4673-6217-7/13/$31.00_c 2013 IEEE

[5]. AmitGupte,Sourabh Joshi, Pratik Gadgul, AkshayKadam, “Comparative Study of Classification Algorithms used in SentimentAnalysis”, International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6261-6264.

[6]. GautamiTripathi and Naganna S, “ Feature Selection And Classification Approach For Sentiment Analysis”,Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.2, June 2015.

[7]. Deepu S. Nair,Jisha P. Jayan, Rajeev R.R, Elizabeth Sherly, “Sentiment Analysis of Malayalam Film Review Using MachineLearning Techniques”, 978-1-4799-8792-4/15/$31.00_c-2015-IEEE In

[8]. Suchita V Wawre1, Sachin N Deshmukh2 , “Sentiment Classification using Machine Learning Techniques”, International Journal of Science and Research (IJSR) Volume 5 Issue 4, April 2016.

[9]. Vishal A. Kharde, S.S. Sonawane , “ Sentiment Analysis of Twitter Data: A Survey of Techniques”,International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016.

[10]. Neha S. Joshi, Suhasini A. Itkat, “A Survey on Feature Level Sentiment Analysis”, International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5422-5425.

[11]. AlessiaD’Andrea, Fernando Ferri, PatriziaGrifoni, TizianaGuzzo ,“Approaches, Tools and Applications for Sentiment Analysis Implementation”,International Journal of Computer Applications (0975 – 8887) Volume 125 – No.3, September 2015.

[12]. Shun Yoshida, Jun Kitazono, Seiichi Ozawa, Takahiro Sugawara, Tatsuya Haga and Shogo Nakamura, “Sentiment Analysis for Various SNS Media Using Naive Bayes Classifier and Its Application to Flaming Detection”,978-1-4799-4540-5/14/$31.00 ©2014 IEEE

[13]. Li Yang, Ying Li , Jin Wang , R. Simon Sherratt, “Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning”, Vol. 8, 2020, IEEE Access, DOI: 10.1109/ACCESS.2020.2969854

[14]. Mahesh B. Shelke, Sachin N. Deshmukh, “Recent Advances in Sentiment Analysis of Indian Languages”, International Journal of Future Generation Communication and Networking, Vol. 13, No. 4, (2020), pp. 1656–1675 ISSN: 2233-7857 IJFGCN

[15]. Pei Ke , HaozheJi , Siyang Liu, Xiaoyan Zhu, MinlieHuangy, “SentiLARE: Sentiment-Aware Language Representation Learning with Linguistic Knowledge”, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 6975–6988, November 16–20, 2020. Association for Computational Linguistics.

[16]. KomangWahyuTrisna& Huang Jin Jie, “Deep Learning Approach for Aspect-Based Sentiment Classification: A Comparative Review”, Applied Artificial Intelligence An International Journal, 36:1, 2014186,2022, DOI: 10.1080/08839514.2021.2014186

[17]. Yuxing Qi, ZahratuShabrina, “Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach”, Social Network Analysis and Mining , 13:31, Springer (2023) https://doi.org/10.1007/s13278-023-01030-x.

[18]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ...&Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

[19]. M. Rathi, A. Malik, D. Varshney, R. Sharma, and S. Mendiratta, "Sentiment Analysis of Tweets Using Machine Learning Approach," 2018 Eleventh International Conference on Contemporary Computing (IC3), Noida, 2018, pp. 1-3.

[20]. I.Iseri, O. F. Atasoy and H. Alcicek, "Sentiment classification of social media data for telecommunication companies in Turkey," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, 2017, pp. 1015-1019

[21]. YAcharya, U. R., Molinari, F., Sree, S. V., Chattopadhyay, S., Ng, K. H., and Suri, J. S., Automated diagnosis of epileptic EEG using entropies. Biomedjgd. Signal Process. Control 7(4):401–408, 2012.

[22]. Imandoust, S. B., &Bolandraftar, M. (2013). Application of k-nearest neighbour (knn) approach for predicting economic events: Theoretical background. International Journal of Engineering Research and Applications, 3(5), 605-610.