International Journal of Scientific Modern Research and Technology
(Vol: 4, Issue: 1, Number: 4)
ISSN: 2582-8150

Classification of Sentimental Reviews using Natural Language Processing Concepts and Machine Learning Techniques

Shailendra Kumar[1] and Dr. Avinash Sharma[2]
[1]PG Scholar, [2]Assistant Professor
[1,2]Dept. of CSE, MITS, Bhopal, India

Abstract- Natural language processing (NLP) is the hypothetically motivated scope of computational strategies for representing and analyzing naturally occurring text at many levels of textual analysis for the goal of attaining automatic language processing system for multiple tasks and applications. One of the most import applications of natural language processing from industry perspective is sentiment analysis. Sentiment analysis is the most eminent branch of NLP because of its capability to classify any textual document to either as positive or negative polarity. With the proliferation of World Wide Web, huge textual unstructured data in form of tweets, messages, articles, social networking discussions, reviews of products and movies are available so as to extract right information from the large pool. Thus, a need is felt to analyze this data to bring out some hidden facts based on the intention of the author of the text. The intention can be either criticism (negative) of product and movie review or it can be admiration (positive). Although, the intention can vary from strongly positive to positive and strongly negative to negative. This thesis completely focuses on classification of movie reviews in either as positive or negative review using machine learning techniques like Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Naive Bayes (NB) classifier. Further, an N-gram Model has been proposed where the documents are classified based on unigram, bigram and trigram composition of words in a sentence. Two dataset are considered for this study; one is a labeled polarity dataset where each movie re- view is either labeled as positive or negative and other one is IMDB movie reviews dataset. Finally, the prediction accuracy of above mentioned machine learning algorithms in different manipulations of same dataset is studied and a comparative analysis has been made for critical examination.
Keyword: KNN, NB, SVM, NLP

I. INTRODUCTION

Natural language processing (NLP) is the investigation of scientific and computational modeling of different parts of language and the advancement of an extensive variety of frameworks. These contain speech recognition frameworks that amalgamate natural language and speech; agreeable interfaces to databases and learning bases that model parts of human-human association; multilingual interfaces; ma- chine interpretation; and message-understanding frameworks. Research in NLP is exceptionally interdisciplinary, including ideas in software engineering, etymology, rationale, and psychology. NLP has an extraordinary part in software engineering on the grounds that numerous parts of the field manage semantic features of reckoning and NLP looks to model language computationally.

At the center of any NLP assignment, there is the critical issue of natural language understanding. The methodology of building computer programs that comprehend natural language includes three noteworthy issues: the first identifies with the point of view, the second addresses to the representation and importance of the phonetic info, and lastly the third and final one address to the world in- formation. Therefore, a NLP framework may start at the word level to focus the morphological structure, nature, (for example part-of-speech). Next, It proceeds onward to the sentence level to focus the word request, language structure, importance of the whole sentence, etc. and afterward to the setting and thegeneral environment or area.

II. PREVIOUS WORK

Pang et.al [2020] have considered sentiment classification taking into account classification perspective with positive and negative sentiments. They have embraced the examination with three different machine learning calculations i.e., Support Vector machine, Maximum Entropy and Naive Bayes classification are being connected over the n-gram techniques.

Turney et. al.[2019] presents unsupervised calculation to order survey as either prescribed i.e., Thumbs up or Thumbs down .The creator has utilized Part of Speech (POS) tagger to distinguish phrases which contain modifiers or intensifiers.

Dave et. al.[2019] had utilized organized survey for testing and training, recognizing features and score strategies to figure out if the reviews are of positive or negative extremity They have utilized the idea of classifier to arrange the sentences retrieved from web search through search crawlers using name of the product asa search quety in crawler program.

Pang and Lee mark [2018] sentences in the report as subjective or goal They have connected machine learning classifier to the subjective gathering, which avoids polarity classification from considering pointless and misdirecting information. They have investigated extraction of strategies on the premise of minimum cut.

Whitelaw et. al [2017] have introduced a sentiment classification technique on the basis of analysis and extraction of appraisal groups Evaluation group corresponds to an arrangement of attribute values in semantic classification.

Li et. al. [2016] have proposed different semi-supervised strategies to tackle the issue of deficiency of marked information for sentiment classification. They utilized sampling technique to manage the issue of sentiment classification i.e., imbalance problem.

Wang and Wang [2015] have proposed a variance mean based feature filtering method that reduces the feature for representational phrase of text classification The performance of the method was observed to be quite comparative as it only considered the best feature and also the computation time got decreased as incoming text was classified automatically.

III. PROBLEM IDENTIFICATION

- Two different approaches of sentiment classification are often used i.e. binary classification of sentiments and multi-class classification of sentiments.
- The work contains a large amount of vague information which is needed to be eliminated.
- The result of each model consists of precision, recall and F-measure as performance evaluation parameters.

IV. RESEARCH OBJECTIVES

The objective of present research is as follows

- supervised machine learning algorithms such as K-nearest Neighbor(KNN), Support Vector machine (SVM) and Naive Bayes (NB)
- These algorithms are implemented on polarity dataset and IMDB dataset and show better result in comparison with the result published in literatures.
- The design of multiband fractal antenna using Koch curve geometry.

V. METHODOLOGY

In this study, labeled polarity movie dataset has been considered which consist of 2000 review, divided equally in to negative and positive reviews and IMDb movie review dataset which consist of 25000 movie reviews for training and same for testing. Each movie review first undergoes through a preprocessing step, where all the vague information is removed. From the cleaned dataset, potential features are extracted. These features are words in the documents and they need to be converted to numerical format.

The vectorization techniques are used to convert textual data to numerical format. Using vectorization, a matrix is created where each column represents a feature and each row represents an individual review. This matrix is used as input to classification algorithm. For

 ISSN: 2582-8150

Polarity dataset, we do not have separate reviews for training and testing so in order to resolve this issue, cross validation technique is applied to choose the training and testing set for each fold. Step-wise presentation of proposed approach is shown in the following block diagram.

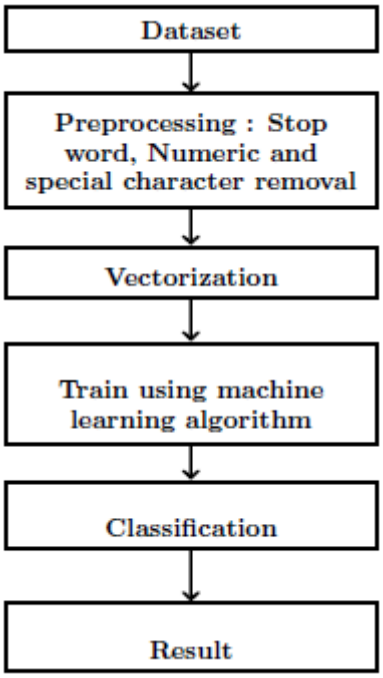

Figure 1: Diagrammatic View of the Proposed Approach

VI. RESULTS AND ANALYSIS

The Precision-Recall Curve of proposed classifier is shown in figure 2. The area of curve computed is 0.91.

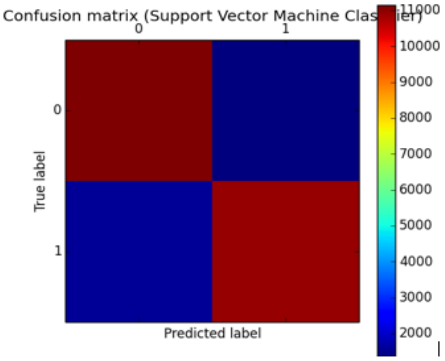

Figure 2: Graphical Presentation of Confusion Matrix for SVM Classifier

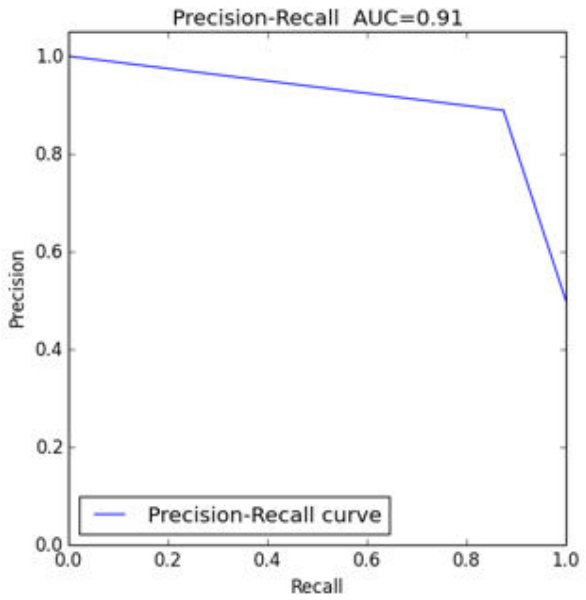

Figure 3: Precision Recall Curve of Proposed Classifier

VII. CONCLUSIONS

This thesis work makes an effort to classify sentiment reviews using supervised machine learning techniques. In this work, three different supervised machine learning algorithms such as K-nearest Neighbor (KNN), Support Vector machine (SVM) and Naive Bayes (NB) are first implemented to check the prediction behavior in classifying the sentimental reviews. Further, using n-gram Model with the application of above mentioned classifying algorithm, the effect of n-gram in classification is studied. These algorithms are implemented on polarity dataset and IMDB dataset and show better result in comparison with the result published in literatures. It is found out that as the value of 'n' in n-gram increases the classification accuracy decreases i.e., for unigram and bigram, The result obtained using the algorithm is remarkably better but when trigram, four-gram, five-gram classification techniques are carried out the accuracy decreases. In future, it is intended to use unsupervised machine learning methods like neural networks and deep learning methods to check the quality of performance. Only three supervised learning methods have been used in this work; so, other supervised learning

 ISSN: 2582-8150

techniques such as Artificial Neural network, Random forest, Decision Tree may also be applied to examine the quality of performance.

REFERENCES

[1]A. K. Joshi, "Natural language processing,"Science, vol. 253, no. 5025, pp. 1242–1249, 1991.

[2]R. Feldman, "Techniques and applications for sentiment analysis," Commu- nications of the ACM, vol. 56, no. 4, pp. 82–89,2013.

[3]M. Mitray, A. Singhalz, and C. Buckleyyy, "Automatic text summarization by paragraph extraction," Compare, vol. 22215, no. 22215, p. 26, 1997.

[4]D. R. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," Computational linguistics, vol. 28, no. 4, pp. 399–408, 2002.

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

[6] IMDb, "Imdb, internet movie database sentiment analysis dataset," 2011.

[7]B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in Proceedings of the ACL, 2004.

[8]A. D. Booth, Machine translation. North-Holland, 1967.

[9]B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classi- fication using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79–86, Association for Computational Linguistics, 2002.

[10]P. D. Turney, "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, pp. 417–424, Association for Computational Linguistics, 2002.