International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

Proposed Model of Hindi Book Review Sentiment Analysis

Raj Firke
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India

Prof. Archana Kadam
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India

***Abstract*—Sentiment analysis, also referred to as opinion mining, is a branch of natural language processing which focuses on the analysis of identifying the opinions or feelings expressed in textual content. The ubiquity of social media platforms and the easy access to enormous volumes of data online have both fueled the present boom in research in opinion mining. Data mining is a process that helps extract useful knowledge from large amounts of data. Analysis of sentiment is performed constantly in widely spoken languages such as English. The amount of scientific work done in regional languages is extremely limited. The primary focus of this study is on conducting a sentiment classification of book reviews that have been written in Hindi, which is a regional language. After the dataset has been acquired, any stop words in it will be eliminated. Next comes TF-IDF and counter vectorization, after which further classification is done based on the algorithm that was employed. Comparison is made between the study of the algorithms that were used in analyzing and the classification of the review using the respective datasets.**

Keywords—TF-IDF, Phrases extraction, fake Review, Regional language.

I. INTRODUCTION

Sentiment analysis, also referred to as opinion mining, is a branch of natural language processing which focuses on the analysis of identifying the opinions or feelings expressed in textual content. The ubiquity of social media platforms and the easy access to enormous volumes of data online have both fuelled the present boom in research in opinion mining. Data mining is a process that helps extract useful knowledge from large amounts of data. Analysis of sentiment is performed constantly in widely spoken languages such as English with Various methods of ML algorithms'. The amount of scientific work done in regional languages is extremely limited. After the dataset has been acquired, stop words in it will be eliminated. Next comes Phrase Extraction in which we concluded that TF-IDF (term frequency–inverse document frequency) provides better and accurate results after which further classification is done based on the Machine Learning algorithms that was employed.

II. RELATED WORK

Less work has been recently done in the area of regional language as there are very less professionals and NLP is a very newly introduced language and very few people have hands-on experience on it.

Parita Shah, et al., "Sentiment Analysis on Movie Reviews in Regional Language Gujarati Using Machine Learning Algorithm," International Journal of Engineering Trends and Technology, vol. 70, no. 3, pp. 319-326, 2022. Major issues in this paper were the gathering of various datasets and efficiently cleaning it for further analysis. The author has performed some preliminary processing on the data in order to obtain the results that were intended to be obtained. As a consequence of this, a list of tokens that can be helpful while performing the task of selecting features has been supplied. The input for multiple machine learning-based classifiers is the feature vector that was generated by utilising the TF-IDF and Count vectorizer techniques respectively. The confusion matrix that is generated by these classifiers is then utilised in order to determine the degree of accuracy possessed by each individual classifier.. As was also noted in this research, there is a possibility of a slight change in accuracy after applying the same model to different datasets; however, the results provided by the suggested model were sufficient. They suggested that in the future, more reviews may be gathered in order to analyse the results obtained when employing the same method on a sizable dataset [1].

Hussaini, et al. (2018). Score-Based Sentiment Analysis of Book Reviews in Hindi Language. International Journal on Natural Language Computing. 7. 115-127. 10.5121/ijnlc.2018.7511. This research investigates the possibility of developing a scored- based opinion mining system for the Hindi language. This system is able to capture the feeling that is conveyed by the words used in book review sentences. The authors carried out three experiments with the help of scores from the Hindi SentiWordNet (H-SWN), in which they began by utilizing the parts-of-speech tags of opinion words in order to derive their potential scores. After then, they concentrated on word sense disambiguation, also known as WSD, in order to achieve a higher.In conclusion, the results of the classification were enhanced by taking into account the morphological variances. The findings were verified by the use of human annotations, which resulted in an accuracy rate of 86.3% overall. Additional work was carried out with the help of the Hindi Subjective Lexicon (HSL). In addition to that, an annotated corpus of Hindi book reviews was produced by them [2].

Kaur, Vipin Deep. "Sentimental analysis of book reviews using unsupervised semantic orientation and supervised machine learning approaches." 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT). IEEE, 2018. The use of sentimental analysis to book reviews is the topic of the research that is presented in this paper. The author has implemented both unsupervised and supervised machine learning techniques, namely NB (Naive Bayes), SVM (Support Vector Machine), and SO-PMI-IR (Semantic Orientation - Pointwise Mutual Information Information Retrieval) approaches, on two publicly accessible book review datasets from Good Reads and Amazon. According to the findings of the comparative study of the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

methods applied to the datasets, the unsupervised method achieves a higher level of performance on the Good Reads dataset, achieving an accuracy of 73.23%. In contrast, the Amazon dataset benefits more from the supervised approach, with Naive Bayes yielding the maximum accuracy (between 73.72% and 74.73%, depending on the number of folds applied to the data) [3]

Sagnika, et al. "A review on multi-lingual sentiment analysis by machine learning methods." Journal of Engineering Science and Technology Review 13.2 (2020): 154. This paper presents the findings of research into multilingual sentiment analysis; it identifies the significant languages that are being acknowledged or that have their own corpus created; it provides a rundown of the many methods that are currently being applied; it describes the contributions of each technique; and it presents the accuracy rates achieved by each technique. Following a discussion of the collection of text or datasets on which sentiment analysis can be performed, the author then enumerated the two primary methodologies that can be utilised for the purpose of conducting sentiment analysis. These methodologies are known as the lexicon based approach and the ML approach. The authors also mentioned the corpus-based and dictionary-based approaches as two other ways in the Lexicon based approach. In addition, supervised, unsupervised, and semi-supervised machine learning techniques can be utilized to carry out the analysis. The authors came to the conclusion that current approaches only have an accuracy rate that is somewhat higher than average, and that higher rates are possible with the discovery of improved procedures and more effective methods [4].

Rohini, V., et al (2016, May). Domain based sentiment analysis in regional Language-Kannada using machine learning algorithm. In 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 503-507). IEEE. In major languages such as English, sentiment analysis is performed on an almost constant basis. The amount of scientific work done in regional languages is extremely limited. This paper compares the results of a study carried out using direct Kannada datasets and one carried out using machine translation in English. The domain-based sentiment analysis of regional language-specific films is the primary emphasis of this particular piece of research. The author did their analysis using a Decision tree-Classifier technique, and they divided the available data set into two parts: test and train. Authors translated the Kannada dataset into English using a machine translation programme. Due to the presence of some ambiguous text in the Kannada dataset, which was not able to be translated to English by a machine, It was found that test data in a regional language produced more accurate findings than English language that had been machine translated. This was a comparison made with the English language. [5] The author of the fifth paper, proposed two methods; one of them which is lexicon based is shown in fig.1 and the other which is machine learning algorithmic based is shown in fig. 2 [5].

Fig. 1- Lexicon based Approach [5]

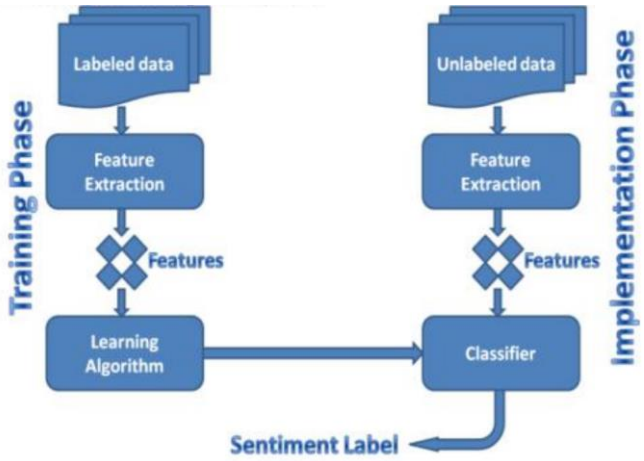

Fig. 2- ML approach [5]

Addanki, et al, Research. (2019). Classification of Book Reviews based on Sentiment analysis: A Survey. 10.13140/RG.2.2.11576.29447. Several different preprocessing techniques, such as the removal of HTML tags and URLs, punctuation, whitespace, removal of special characters, and stemming, are presented in this paper. The paper also discusses how machine learning algorithms will be used to perform opining mining for classification of reviews in order to recommend specific books based on user interest factors. These techniques are used to remove noise, and the paper also discusses how machine learning algorithms will be used. This paper gives an overview of the many algorithms that are used in the sentiment analysis book recommendation system. In addition to this, it analyses and contrasts the majority of the characteristics shared by the various algorithms. The various methods that can be used to improve this classification and clustering process in terms of time, accuracy, scalability, and overall performance are expounded on and explored below. The authors used datasets that are freely available from Library Thing, Amazon reviews, and INEX Book Track, and after completing analysis, they gave accuracy of models and found the optimal method for doing sentiment analysis on book reviews. [6] The author elaborated the techniques available for performing sentiment analysis and drew a chart for the following.

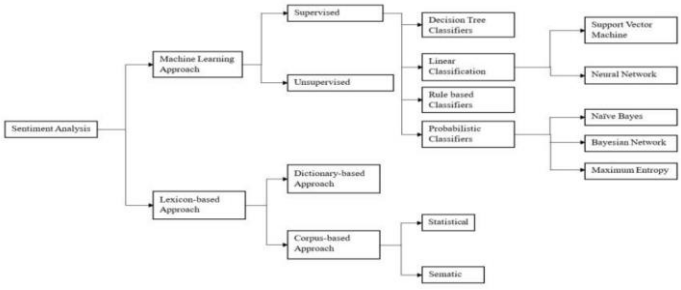

Fig. 3– Sentiment Classification techniques [6]

Referring to above Fig. 3, it is seen that the methodology proposed by the authors consists of 2 main methods i.e. Machine Learning and lexicon based approaches which are further subdivided into other methods.[6]

Common findings from Literature review:

· Sentiment analysis is done by using 2 methods i.e. lexicon based and machine learning based.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

· Data preprocessing is a very important step before performing analysis on the data collected or gathered.

· Accuracy of the models are above average and there is a lot of work needed to be done in the regional language area.

III. PROPOSED MODEL

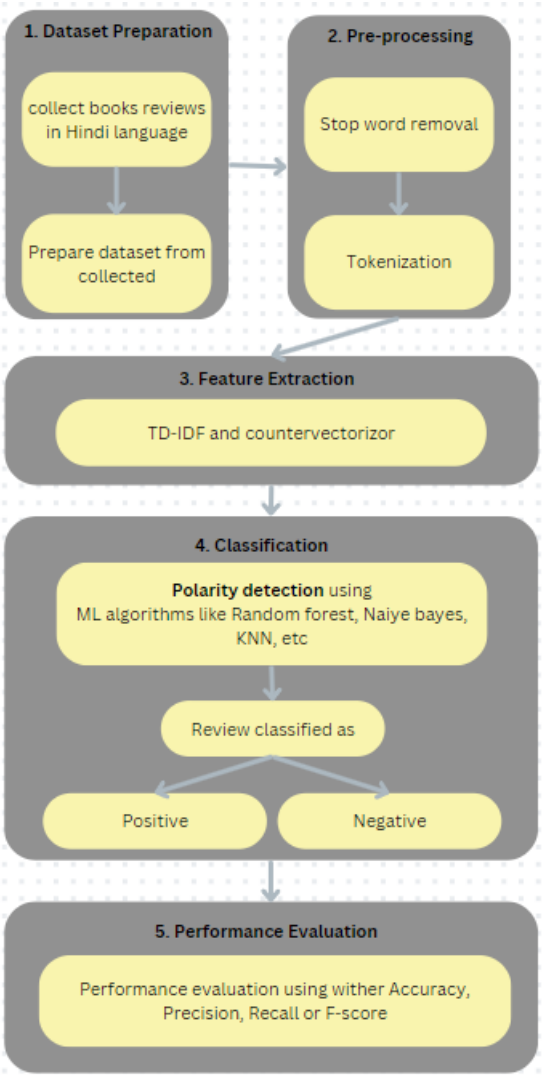

Fig. 4- Simplified Stage-wise Process Diagram

After studying Related work, and searching common findings from the surveys, we have tried to propose a system which will give sentiment analysis on the basis of a machine learning model.

The technology and data stream methodology for the research that is being suggested may be seen depicted in the following figure. The procedure can be broken down into the following phases for easier comprehension:

Step 1: Preparation of dataset –

The primary aim of this step is to collect data for analysis and then creating a dataset from it. In our situation the data is book reviews in Hindi Language. We will be gathering all reviews from the website or blogs based on book review in Hindi Language. After collecting it we will make a dataset on which Sentiment Analysis will be performed.

Step 2: Pre-processing of dataset-

From step 1, we get dataset ready to work on but before performing analysis; we need to make sure it is clean and perform appropriate operation to make it usable for us. The act of preparing the raw data and making it appropriate for use in a machine learning model is referred to as "data preparation." It is the initial and most important phase in the process of developing a machine learning model. We will be cleaning the data using Data prepressing techniques and further we will remove the stop words. Stop words are words in a stop list that are filtered out either before or after the processing of natural language data due to the fact that they do not contribute significantly. We further perform Tokenization on them using NLTK Library. The Natural Language Toolkit (NLTK) is a framework for developing Python applications that utilise human language data for statistical natural language processing (NLP). It includes libraries for tokenization, parsing, classification, stemming, tagging, and semantic reasoning.

Step 3: Extraction of Features-

Feature extraction is a feature of the dimensionality reduction procedure, in which an initial collection of unprocessed data is partitioned into more manageable categories. Therefore, processing will be simplified. We use TF-IDF; In information retrieval, tf–idf, an abbreviation for term frequency–inverse document frequency, is a numerical statistic meant to reflect a word's significance to a document in a collection or corpus. It is being utilised as a weighing factor. We execute feature extraction as it enables the model to be constructed with less machine effort and accelerates the machine learning and generalisation processes.

Step 4: Classification-

After the data has been pre-processed and its features extracted, we use machine learning algorithms to identify the polarity of the data and then classify it as either positive or negative depending on that polarity. We make use of various classification methods, such as rain forest, KNN, and Naive Bayes, among others, in order to categorise the reviews according to the words used.

Step 5: Performance Evaluation-

When it comes to developing an efficient machine learning model, one of the most crucial phases is to conduct an evaluation of the performance of the machine learning model. Different metrics are utilised, and these metrics are classified as performance metrics or evaluation metrics depending on whether they are used to evaluate the quality of the model or its performance. After training the model, we will test the model and use performance evaluation techniques like Accuracy, Recall or F-score.

IV. METHODS

A. Random Forest-

It is a computation for directed learning, and depending on your requirements, you can make use of it either for the purpose. It is the computation that offers the greatest degree of adaptability and is the easiest to use. Trees are the fundamental building blocks of a forest. It is believed that a timberland's ability to build choice trees based on arbitrarily selected Sentiment Analysis on Book Reviews in Regional Language Hindi `Department of Computer Engineering, PCCOE, Pune

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

Page 19 informational collections, make predictions based on the knowledge gleaned from each tree, and select the approach that yields the best results can be directly correlated to the number of trees present. In addition to this, it is able to produce a respectable profit based on the significance of the ability.

B. Multinomial Naïve Bayes-

This algorithm is a probabilistic deep learning that is utilised in Natural Language Processing the majority of the time (NLP). The Bayes theorem serves as the foundation for the computer programme that can determine the tag associated with a piece of text such as an email or a newspaper article. After determining the probabilities of having each tag for a specific sample, it selects the tag with the highest likelihood as the one to output.

C. K-nearest neighbour-

The KNN algorithm adheres to the principle of similarity by determining the distance (in Euclidean units) between different focuses. It is necessary to determine the order of 16 things based on the distance initially. Following that, it will make an effort to predict information concentrates that are located.

D. Support Vector Machine-

The goal of this calculation is to locate a hyperplane in N-dimensional space that is capable of independently grouping the information focuses (N-number of attributes). There are a few different hyperplanes that may be picked to recognise the two gathering centers, and one of these options is available to you. Our goal is to locate a plane that offers the most advantage, such as the greatest distance between the information centers of the two different categories. The expansion of the hole from the edge provides some assistance in order to achieve the purpose of arranging additional certainty in anticipated information focuses.

V. CONCLUSION

Analysis of sentiment is performed constantly in widely spoken languages such as English with Various methods of ML algorithms'. The amount of scientific work done in regional languages is extremely limited. After analysing various aspects of people in sentiment analysis, we were able to conclude that if we convert Hindi into English using machine translation and perform analysis on it; still the analysis done directly in Hindi language rather than translating it to English is be more accurate and precise. We also learned various approaches to perform sentiment analysis and how to break language barriers. We were able to conclude that Machine learning needs more or bulk dataset so that it is trained properly and we can test it Whereas lexicon based approach consist of dictionary which contains positive and negative words and on the basis on that it is performed and it will give better results on smaller datasets. Also in machine learning, It's possible that the unsupervised learning model will produce less accurate results than the supervised learning model

REFERENCES

[1] Parita Shah, Priya Swaminarayan, Maitri Patel, Nimisha Patel, "Sentiment Analysis on Movie Reviews in Regional Language Gujarati Using Machine Learning Algorithm," International Journal of Engineering Trends and Technology , vol. 70, no. 3, pp. 319-326, 2022.

[2] Hussaini, Firdous & Savaram, Padmaja & S, Sameen. (2018). Score-Based Sentiment Analysis of Book Reviews in Hindi Language. International Journal on Natural Language Computing. 7. 115-127. 10.5121/ijnlc.2018.7511

[3] Kaur, Vipin Deep. "Sentimental analysis of book reviews using unsupervised semantic orientation and supervised machine learning approaches." 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT). IEEE, 2018

[4] Sagnika, Santwana, et al. "A review on multi-lingual sentiment analysis by machine learning methods." Journal of Engineering Science and Technology Review 13.2 (2020): 154

[5] Rohini, V., Thomas, M., & Latha, C. A. (2016, May). Domain based sentiment analysis in regional Language-Kannada using machine learning algorithm. In 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 503-507). IEEE

[6] Addanki, Mounika & Saraswathi, Dr & Scholar, Research. (2019). Addanki, Classification of Book Reviews based on Sentiment analysis: A Survey. 10.13140/RG.2.2.11576.29447