International Journal of Engineering Innovations in Advanced Technology

ISSN: 2582-1431 (Online), Volume-4 Issue-4, December 2022

IMDB Movie Review Analysis Using Traditional Machine Learning Models

Kakarla Phani Sravya

Research scholar (CSE) in International School of Technology and Sciences for Womens (ISTS women's college) in NH-16, East Gonagudem, Rajanagaram, Andhra Pradesh, 533294

.E-mail: Sravyakakarla777@gmail.com

Jamanania

Associate professor and hod (CSE) in International School of Technology and Sciences for Womens (ISTS women's college) in NH-16, East Gonagudem, Rajanagaram, Andhra Pradesh, 533294

Abstract: Sentiment analysis is the analysis of emotions and opinions from any form of text. Sentiment analysis is also termed as opinion mining. Sentiment analysis of the data is very useful to express the opinion of the mass or group or any individual. This technique is used to find the sentiment of the person with respect to a given source of content. Social media and other online platforms contain a huge amount of the data in the form of tweets, blogs, and updates on the status, posts, etc. In this paper, we have analyzed the Movie reviews using various techniques like Naïve Bayes, K-Nearest Neighbor and Random Forest.

KEYWORD: opinion mining, **source of content**, **Social media** , **online platforms** ,**Movie reviews.**

I. INTRODUCTION

With the advent of Web 2.0 various platforms like Facebook, Twitter, LinkedIn, Instagram allows citizens to share their comments, views, feelings, judgements on the myriad of topics ranging from education to entertainment. These
platforms contain the huge amount of the data in the form of tweets, blogs, and updates on the status, posts, etc. Sentiment Analysis aims to determine the polarity of emotions like happiness, sorrow, grief, hatred, anger and affection and opinions from the text, reviews, posts which are available online on these platforms. Opinion Mining finds the sentiment of the text with respect to a given source of content. Sentiment analysis is complicated because of the slang words, misspellings, short forms, repeated characters, use of regional language and new upcoming emoticons. So, it is a significant task to identify appropriate sentiment of each word. Sentiment Analysis is one of the most active research areas and is also widely studied in data mining. Sentiment analysis is applied in almost every business and social domain because opinions are central to most human activities & behaviors. Sentiment analysis is very popular because of its efficiency. Thousands of documents can be processed for sentiment analysis. Since it is an efficient process which provides good accuracy, therefore it has various applications:

1. Purchasing Merchandise or Service: While purchasing a merchandise or service we must take a right decision which is not a difficult task anymore. By sentiment analysis, people can easily evaluate reviews and opinions of any commodity or service and can effortlessly
compare the competing brands.
2. Quality Improvement in Product or Service: By Opinion mining, the producers can collect the user's opinion whether favorable or not about their product or service and then they can enhance and upgrade the quality of their product or service.
3. Recommendation Systems: By analyzing and categorizing the people's opinion according to their preferences and interests, the system can predict which item should be recommended and which one should not be recommended.
4. Decision Making: People's sentiments, ideas, feelings are very important factor to make a decision. While buying any item be it book or clothes or electronic items user's first to read the opinions and reviews of that particular product and those reviews have a great impact
on user's mind.
5. Marketing research: The result of sentiment analysis techniques can be utilized in marketing research. By this technique, the attitude of consumers about some product or services or any new government policy can be analyzed.

ISSN: 2582-1431 (Online), Volume-4 Issue-4, December 2022

6. Detection of flame: The monitoring of newsgroups, blogs and social media is easily possible by sentiment analysis. This technique can detect insolent, arrogant, over heated words used in tweets, posts or forums and blogs on the internet. There are following phases of Sentiment Analysis:

Pre-Processing Phase: The data is first cleaned to reduce noise.

Feature Extraction: A token is given to the keywords and this token is now put under analysis.

Classification Phase: Based on different algorithms these keywords are put under certain category.

II. LITERATURE SURVEY

Joscha et. al, in their paper devised and compared various techniques like Bag of words models, n-grams for using semantic information to improve the performance of sentiment analysis. The earlier approaches did not consider the semantic associations between sentences or documents parts. Research by A. Hogenboom et al. neither compared the methodological variants nor provided a method to merge disclosure units in the most favorable manner. They aimed to improve the sentiment analysis by using Rhetoric Structure Theory (RST) as it gives a hierarchical representation at the document level. They proposed an integration of the grid search and weighting to find out the average scores of sentiments from Rhetoric Structure Theory (RST) tree. They encoded the binary data into the random forest by using feature engineering as it greatly reduced the complexity of original RST tree. They concluded that machine learning raised the balanced accuracy and gives a high F1 score of 71.9%. Amir Hossein Yazdavar et al. in this paper [3] provided novel understanding of sentiment analysis problem containing numerated data in drug reviews. They analyzed sentences which contained quantitative terms to classify them into opinionated or non-opinionated and also to identify the polarity expressed by using fuzzy set theory. The development of fuzzy knowledge base was done by interviewing several doctors from various medical centers. Although the number of researches has been done in this field (Bhatia, et al., [4]) these do not consider the numerical (quantitative) data contained in the reviews while recognizing the sentiment polarity. Also, the training data used has a high domain dependency and hence cannot be used in different domains. They concluded that their proposed method knowledge engineering based on fuzzy sets was much simpler, efficient and has high accuracy of over 72% F1 value. Dhiraj Murthy in his paper [5] he identified what roles do tweets play in political elections. He pointed out that even though there were various researches and studies done to find out the political engagement of Twitter, no work was done to find out if these tweets were Predictive or Reactive. In his paper, he concluded that the tweets are more reactive than predictive. He found out that electoral success in not at all related to the success on Twitter and that various social media platforms were used to increase the popularity of a candidate by generating a buzz around them. Ahmad Kamal in his paper [6] designed an opinion mining framework that facilitates objectivity or subjectivity analysis, feature extraction and review summarization etc. He used supervised machine learning approach for subjectivity and objectivity classification of reviews. The various techniques used by him were Naive Bayes, Decision Tree, Multilayer Perceptron and Bagging. He also improved mining performance by preventing irrelevant extraction and noise as in Kamal's paper. [7]. Humera Shaziya et al. in this paper [8] classified movie reviews for sentiment analysis using WEKA Tool. They enhanced the earlier work done in sentiment categorization which analyzes opinions which express either positive or negative sentiment. In this paper, they also considered the fact that reviews that have opinions from more than one person and a single review may express both the positive and negative sentiment. They conducted their experiment on WEKA and concluded that Naïve Bayes performs much better than SVM for movie reviews as well as text. Naive Bayes has an accuracy of 85.1%. Akshay Amolik et. al. in his paper [9] created the dataset using twitter posts of movie reviews and related tweets about those movies. Sentence level sentiment analysis is performed on these tweets. It is done in three phases. Firstly, preprocessing is done. Then Feature vector is created using relevant features. Finally, by using different classifiers like Naïve Bayes, Support vector machine, Ensemble classifier, kmeans and Artificial Neural Networks, tweets were classified into positive, negative and neutral classes. The results show that we get 75 % accuracy form SVM. He negated Wu et. al.

ISSN: 2582-1431 (Online), Volume-4 Issue-4, December 2022

paper [10] which made an observation that if @username is found in a tweet, it influences an action and also helps to influence the probability. But in this paper Akshay Amolik
replaced @username with AT_USER and hashtags were also removed due to which we used Support Vector Machine rather than Naive Bayes which increased the accuracy by
10%.

III. MACHINE LEARNING METHODS

3.1 Naïve Bayes

It is a technique based on Bayes' Theorem. Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. This model is easy to build and particularly useful for very large datasets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

$$P(C|X) = P(X|C) * P(C)/P(X) \quad (1)$$

P(C|X) is posterior probability of class C
P(C) is prior probability of class C
P(X|C) is probability of predictor given the class. P(X) is prior probability of predictor 3.2 K- Nearest Neighbor K-NN is the simplest of all machine learning algorithms. The principle behind this method is to find a predefined number of training samples closest in distance to the new point and predict the label from these. The number of samples can be a user-defined constant or vary based on the local density of points. The distance can be any metric measure. Standard Euclidean distance is the most common choice for calculating the distance between two points. The Nearest Neighbors have been successful in a large number of classification and regression problems, including handwritten digits or satellite image processing and so on.

3.3 Random Forest

Random Forests are the learning method for classification and regression. It constructs a number of decision trees at training time. To classify new case, it sends the new case to each of the trees. Each tree perform classification and output a class. The output class is chosen based on majority voting that is the maximum number of similar classes generated by various trees is considered as the output of the Random Forest.

Random Forests are easy to learn and use for both professionals and laypeople with little research and programming required. It can easily be used by persons that don't have a strong statistical background.

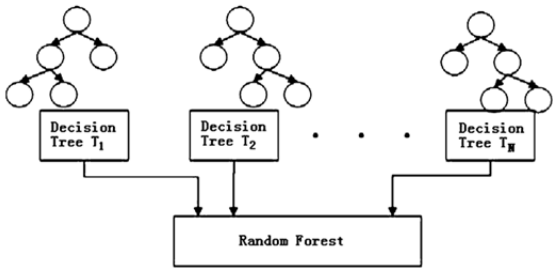

Fig 1: Random Forest figure taken from techleer.com

IV. EXISTINGSYSTEM

In existing system could not work effectively. In existing systems, the large amount of data maintenance is difficult and could not give the accurate results also.

Disadvantages of Existing System

1. Less accuracy.
2. Data maintenance is difficult

V. PROPOSED SYSTEM

In this paper, movie reviews are classified into positive or negative polarity. The system can be used to classify a huge database of movie reviews. Best thing about the system that it is a web-based API for sentiment analysis for movie reviews with JSON output to display results on any operating system.

Advantages of Proposed System:

1. Easy to interact.
2. Accuracy is more.

VI. MODULES

1. Upload IMDB Movie Review Dataset
2. Load Sentiment Classifier
3. Identify Sentiment Polarity
4. Sentiment Analysis Graph

1) Upload IMDB Movie Review Dataset: In this module user can upload reviews from any domain such as movie review, product reviews etc.

2) Load Sentiment Classifier: In this module sentiment classifier will be loaded and this contains Naïve Bayes classifier implementation which can identify POS (parts of speech) from words and then calculate polarity from the given reviews. We are

ISSN: 2582-1431 (Online), Volume-4 Issue-4, December 2022

using python inbuilt sentiment classifier called 'SentimentIntensityAnalyzer'.

3) Identify Sentiment Polarity: After calculating polarity we can use this module to identify polarity of each word and if more number of words give high polarity for POSITIVE then that sentence will be consider as positive and if more number of words give negative polarity then that sentence consider as negative.

4) Sentiment Analysis Graph: This module display graph which shows total reviews found and how many are positive reviews and how many are negative reviews identified from entire dataset.

VII. CONCLUSION

In this paper, movie reviews are classified into positive or negative polarity. The system proposed by author in the paper can be used to classify a huge database of movie reviews. Best thing about the system that it is a web-based API for sentiment analysis for movie reviews with JSON output to display results on any operating system. Table 1 shows that the system works decently. This will help movie producers to check the status of their movie. Future work, this API can be trained for other reviews like smartphones, laptops or clothes etc.

VIII. REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135.

[2] M. Hu and B. Liu, "Mining and summarizing customer reviews," Proceedings of the tenth ACM international conference on Knowledge discovery and data mining, Seattle, 2004, pp. 168-177.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol.10, 2002, pp. 79-86.

[4] Jie Yang University of Wollongong, Australia "Mining Chinese social media UGC- a big-data framework for Extract Reviews POS Tagging Opinion Word List Identify Sentence Polarity Count & Temporary Save Sentence Polarity Check If Any Sentence is Left For Analysis ? Count Total Positive and Negative Polarity of all Sentences Review Classification and Generate Results N o Yes Identify Review Polarity International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 16 (2018) pp. 12788-12791 © Research India Publications. http://www.ripublication.com 12791 analyzing Douban movie reviews", Journal of Big Data Springer, 2016

[5] Kia Dashtipour Scotland, United Kingdom "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques", Springer, 2016

[6] Kigon Lyu Korea University, Korea "Sentiment Analysis Using Word Polarity of Social Media", Springer, 2016

[7] Monu Kumar Thapar University, Patiala "Analyzing Twitter sentiments through big data", IEEE, 2016

[8] Minhoe Hur Seoul National University "Box-office forecasting based on sentiments of movie reviews and Independent subspace method", Information Sciences, 2016

AUTHORS PROFILE

Kakarla Phani Sravya completed her B.Tech in 2019 in Computer Science and Engineering and has interest in DBMS ,Operating systems, Software Testing, Machine Learning .Sentiment Analysis of movie review using supervised machine learning as a part of research.

Mr. Jamnania (M.Tech, Phd) he working as an associate professor in ISTS women's engineering college