e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:04/Issue:06/June-2022 Impact Factor- 6.752 www.irjmets.com

SENTIMENT ANALYSIS OF MOVIE REVIEWS USING SUPERVISED MACHINE LEARNING TECHNIQUES

Gogi Sai Chaitanya*1, Ryakala Pravallika*2, Katravath Priyanka*3, Neha Kumari*4, M. Rajkumar*5

*1,2,3,4B.Tech Student, Department Of Computer Science And Engineering, JB Institute Of Engineering And Technology, Hyderabad, Telangana, India.

*5Assistant Professor, Department Of Computer Science And Engineering, JB Institute Of Engineering And Technology, Hyderabad, Telangana, India.

ABSTRACT

Sentiment Analysis is a new concept in Research and is useful in many other fields. Most of the data is collected from surveys, comments, and reviews. All the collected data is used to improve products and services provided by both private organizations and governments around the world. In this project sentiment analysis of reviews can be done using opinion mining, feature based and supervised machine learning. The objective of the paper is to determine polarity of reviews using parts of speech i.e., nouns, verbs, and adjectives as opinion words. Here, the reviews are classified into positive and negative categories. Reviews of IMdb are used as source dataset and NLTK for parts of speech tagging. This project also contains some facts about the classification of data on basis of polarity.

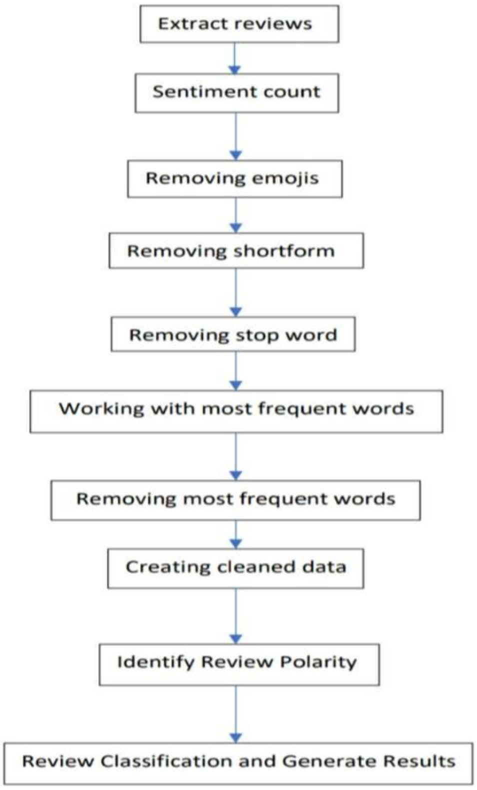
I. INTRODUCTION

Everyone makes decisions based on their understanding, sentiments or opinion passed by other persons. Whenever an individual wants to buy a product, they look for the opinions from others about the item or product. Similarly, every firm wants to give their best product to the market so they collect opinion from their customers about their product using surveys. Sentiment Analysis is a study of opinions, sentiments or emotions conveyed about a product or a movie. The advancement in the field of web technology has changed the way in which people can express their views. For the analysis of products while shopping online or while booking movie tickets for watching movies in theatres, people depend upon this user created data. The users can interact through posts, Facebook, tweets, and hash tags. The amount of data is so huge that it is difficult for a normal human to analyze and conclude. Sentiment analysis is concerned with the identification and classification of opinions. It is broadly classified in the two types first one is a knowledge-based approach and the other machine learning techniques. For identifying opinions firstly we need a large datasets of emotions and effective knowledge. On the other hand, the Machine learning approach makes use of a training data set and a test data set to develop a classifier. It is simpler than Knowledge base approach. While developing the algorithms we face several obstacles in Sentiment analysis. The first is that an opinion word can be positive or negative depending upon the situation. Next conflict is that everyone does not always convey their opinions in the comparable way. Opinion mining helps to understand the relationship between textual reviews and the consequences of those reviews.

Objective:

Our important objective is to classify movie reviews into positive or negative polarity by using supervised machine learning. Now a days the experience or opinion passed by others is used by human beings to make their decisions. The objective of this paper is to recommend the best movie based on polarity of reviews. In modern world there are many movies are available to watch but the best is recommended using the polarity of reviews to the users. As many movies are available to watch it makes user difficult to choose a movie to base on their interest. The main objective is to obtain the polarity of reviews. The proposed system would help to minimize the difficulties faced by users in choosing a movie and maximize the service provided by the organisations from the collected data.

System Architecture:



II. METHODOLOGY

Dataset: The dataset used for this task was collected from Keras. The dataset contains 50,000 training examples collected from IMDb where each review is labelled with the rating of the movie on scale of 1-10. As sentiments are usually bipolar like good/bad or happy/sad or like/dislike, we categorized these ratings as either 1 (like) or 0 (dislike) based on the ratings. If the rating was above 5, we deduced that the person liked the movie otherwise he did not. Primarily the dataset was divided as training and testing subsets containing 25,000 reviews in each subset. Due to the number of training examples were small and leading to unfit we came to know. We then tried to redistribute the examples as 40,000 for training and 10,000 for testing. While this resulted good models, it may lead to unfit on reviews and bad result on the test set.

Extraction: We used 3 methods for extraction of meaningful features from the review text which could be used for training purposes. These features were then used for training several classifiers.

Bag of Words: In any text mining process, this is the typical was for representation of word. Here total word counts are calculated for each word for all the reviews and then this data is used to make different feature representationsews and then this data is used to make different feature representations. As the total number of words in the dictionary was the first feature set was created using only the 50,000 most frequent words according to their occurrence. Accross the whole dataset, using all the words that occured at least twice are used to create the another representation of bag of words to the previous representation. Then we can make sure that we delete the most mistyped words. And also if the words are occured only once in the dataset would put up nothing to the classifier. Another feature representation was created along the same lines but with words occurring at least 5 times. The size was 34,000 and 76,000, respectively for those two features representations.

III. MODELING AND ANALYSIS

Exploratory Analysis: While working with review text first we need to estimate the average size of reviews. Every review is differ from another one so these are represented through graphs. From this information we noticed that in general people tend to write brief reviews for movies and as such this is a good topic for sentiment analysis. Also, people write reviews when they have strong opinions about a movie; they either loved it or hated it. Not only word count from reviews but we can also estimate occurances of the words of reviews.

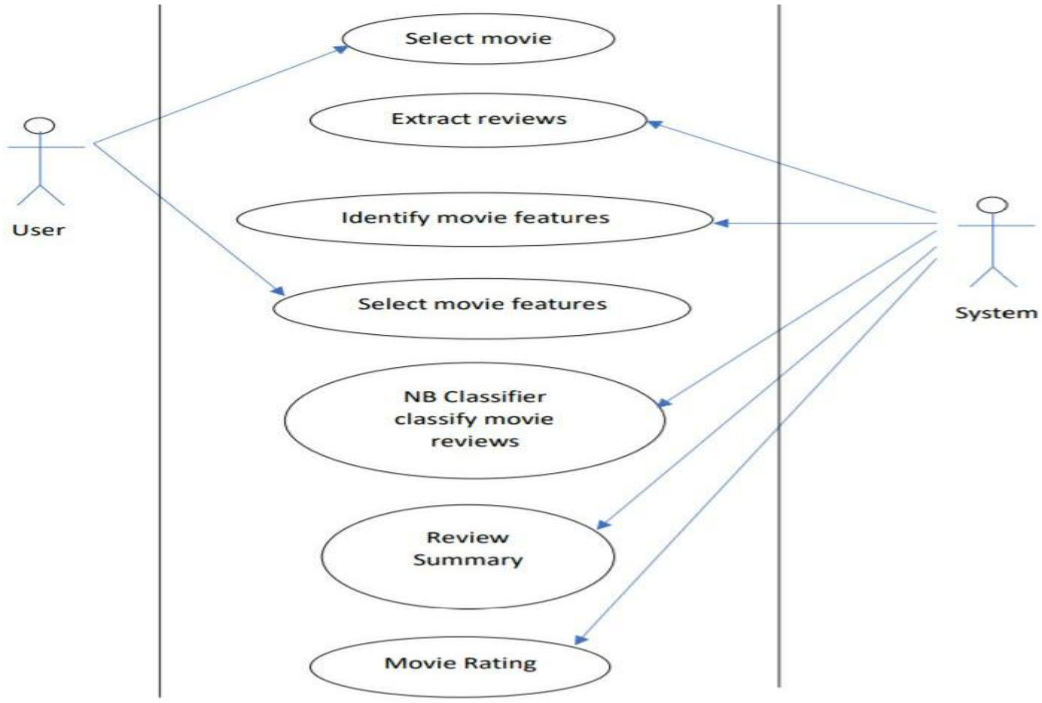
Due to their relative importance some words may occur more than others. Here we listed 20 most occurring words both in negative and positive reviews along with a graph showing variability of word occurrences across all reviews. Also, the average word occurrence count was around 33 over all 50000 reviews. From all the collected data or sources and the below graphs, it is clear that "Bag of Words" is not an exceptionally good model for doing sentiment analysis of reviews because similar words have high counts in both positive and negative reviews. Also, overall number of unique words is massive across all the reviews and hence we use only top 50,000 and 1,00,000 of these during training. Because of this we moved on to other methods of feature extraction like n-gram modelling and TF-IDF counts of each word.

N-Gram Modelling: Here the bag of words concept neglects the semantic context of the reviews and focuses mainly on frequency of each word. To overcome that, we also tried ngram modelling wherein we created unigrams, bigrams, and mixture of both. While creating unigrams is like the bag of words approach, bigrams provided more contextual information on the review text. We created one feature representation like the "Bag of Words" approach above but using the bigrams. Also, to get more insight on textual information we created a feature set using a mixture of n-gram with n = 5 and using only those grams with minimum count of 10. In case of n-gram modelling, we did not remove the stop words as we were doing for previous cases.

TF-IDF Modelling: While the two methods of feature extraction descried above concentrated more on higher frequency parts of the review, they completely ignored the portions which might be less frequent but have more significance for the overall polarity of the review. The feature representation for this model is like the Bag of Words model except that we used TF-IDF values for each word instead of their frequency counts. To limit the number of words common to both positive and negative reviews, we ignored all the words whose count was more than 50 as they would not contribute much to the classifier.

Models: Classification of reviews as favourable or unfavourable is the overall task of this project. Therefore, for this classification task we explored multiple classification models on above feature representations. We also used other naïve classification models like Complement NB Model, Multinomial NB Model and Bernoulli NB Model. Apart from these, we also trained the above feature representations on Naïve Bayes' Classifier as this is primarily used in case of text mining in combination with Bag of Words and N-Gram Modelling. We also trained a model to match the similarity between the reviews and classify them accordingly. For all the above models, we used sklearn modules by tuning their parameters and not changing their implementations and so we will not go into their theory in this report.

Use Case Diagram:



IV. RESULTS AND DISCUSSION

Data Preparation: Data has been collected from the social networking websites. We have extracted the data of a particular movie in a single text document file as we were working on the document level sentiment classification.

EDA and Data Pre-Processing: Pre-processing of data has been done because the data we extracted were in raw form which was not suitable for the sentiment analysis to provide better accuracy. Since we have stored the data in text format, our analysis tool has failed to read in text format. However, Weka has provided the inbuilt converter such as Text Directory Loader which will load the dataset in text directory format.

Creating Dataset: The dataset used for this task was collected from Keras. The dataset contains 50,000 training examples collected from IMDb where each review is labelled with the rating of the movie on scale of 1-10.

Applying Algorithm: It is a technique based on Bayes' Theorem. The various Naive Bayes classifiers assumes that the presence of a feature isn't affected by another feature. This model is easy to build and particularly useful for large datasets. Naive Bayes is known for its simplicity and it also outperforms even complex classification methods

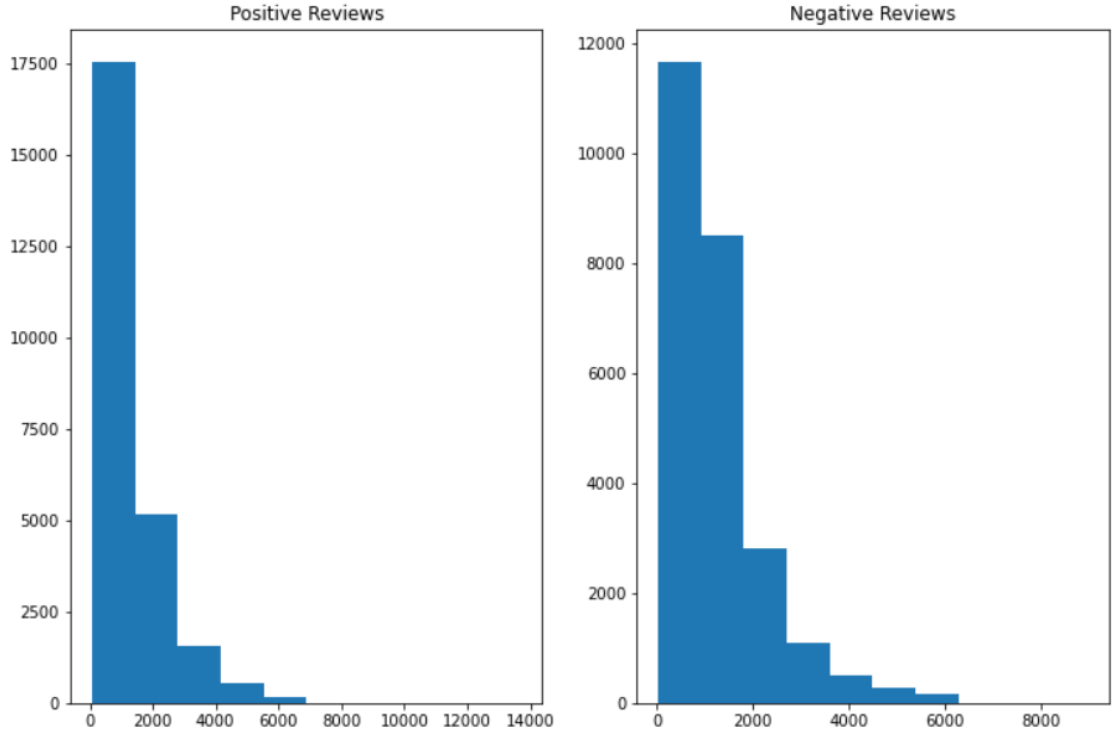


Figure: Output Screens

V. CONCLUSION

In this research, various NB techniques were used to identify the polarity of the movie reviews. The algorithms performed were Complement NB Model, Multinomial NB Model, Bernoulli NB Model. The Complement NB Model achieved 86.35% accuracy, Multinomial NB Model achieved 86.35%accuracy, Bernoulli NB Model achieved 85.05% accuracy. Finding the polarity of the reviews can help various organizations and in various domain. By creating the intelligent systems we can provide the users with comprehensive reviews of movies, products, services etc, without requiring the user to go through individual reviews. By this he can directly take decisions based on the results provided by the intelligent systems.

VI. 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 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.

[9] Jorge A Balazs University of Chile “Opinion Mining and Information Fusion- A survey”, 2015.

[10] Donglin Cao Xiamen University, China “A cross-media public sentiment analysis system for microblog”, Springer, 2014.