TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 4427-4431

Sentiment Analysis on Movie Reviews

[1]P. Trupthi, [2]Pandeti Madhura, [3]Pooja Jehan, [4]Pooja Polampalli, [5]Aruna Kumara B

[1,2,3,4]School of C & IT, REVA University, Bengaluru, India
[5]Assistant Professor, School of C & IT, REVA University, Bengaluru, India
1trupthiputhamakula@gmail.com, [2]madhurapandeti@gmail.com, [3]poohbear.pooj@gmail.com, [4]poojapolampalli14@gmail.com, [5]arunakumara.b@reva.edu.in

Article Info
Volume 83
Page Number: 4427-4431
Publication Issue:
May - June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract

Sentimental analysis of movie reviews is basically excavating the judgements based on the existing movie analysis on the internet. This approach identifies and categorizes the subjective opinions to decide the attitude of the subject towards a particular text. These subjective opinions determine the satisfaction level of the subject. These subjective opinions also determine the success or the failure rate of the text. The polarity or the differences between these opinions are found using various pre-processing techniques i.e. filtering of data by converting the text into words, removing the stop words and negating the neutral opinions. And for the faster processing, Feature Extraction and Feature Selection techniques are used as it reduces the word count by eliminating redundant contents. The accuracy of these subjective opinions is governed by using certain classifiers such as Naive Bayes, Logistic Regression, Decision Tree, Random Forests, KNN, SVM, and Adaptive Boosting. Each classifier has a different accuracy rate processes based on the given code values. This paper focuses to find out which classifier gives the most accurate result.

Keywords**:** *Sentiment Analysis, Polarity, Sentiments, Movie reviews, Feature extraction, Classifier, Feature selection, Accuracy.*

1. Introduction

The internet plays a vital role in everyone's life. Everyone depends on the internet for everything. Even to buy a product, people go for online shopping. To watch a movie, now, the internet provides various platforms. But to know the standard of a movie, people depend on the subjective reviews of the movie which provides emotional tranquillity from an individual perspective. These reviews give people an insight into the movie depending on the polarity of it. When we rate a movie by providing stars, which will give us an idea about the success and failure of a movie. A collection of movie reviews gives us an in-depth view of different aspects of the movie. Textual movie reviews give us the strong and weak points of the movie, but the accuracy of these subjective valuations is unknown to the reader. Therefore, certain measures are used to find the accuracy of these subjective reviews which would help the reader. This is the sentiment analysis of the reviewer which can help the reader analyse the state of mind of the reviewer, that is, if the reviewer felt "happy", "sad", or "angry" or "satisfied or dissatisfied" with the movie.

The aim is to predict the overall polarity of the movie by making use of the relationship between the words used in the subjective reviews. Predicting the polarity will help in analysing the sentiment of the text used and to categorize the text under a specific class or category. To categorize the text, certain pre-defined classifiers are used which will help in polarizing the text to either positive or negative. The neutral categorization is omitted as it is considered invalid in this paper. More classifiers, other than the pre-defined classifiers, can also be used for the same, which has to be defined afresh.

The paper is ordered as follows: Section 2 reviews the analysis of various machine learning techniques implemented in sentiment analysis of movie reviews.

May-June 2020
ISSN: 0193-4120 Page No. 4427-4431

Section 3 gives the approach for sentiment analysis and classification of movie reviews. Section 4sets forth the result of the proceeding. In the end, in Section 5, the conclusion of the work is presented.

2. Related work

Tirath Prasad Sabu and Sanjeev Ahuja of NIT, Raipur made a study on the movie reviews online and how certain use of sentiments in the text affect the polarity of the film industry. This paper suggests that sentiment analysis can help us determine the tone of review of movies using the words the reviewers use on social media platforms to classify if the movie is positive, negative or neutral in public opinion. The suggested use of different algorithms like feature selection, Information gain and SentiWordNet which follow a lexical approach that analyse the use of grammatical pattern and the use of certain words in movie reviews and how those affect the sentiments of the audience. According to the paper, opinion mining from movie reviews which are structured, semi-structured or unstructured, is done with the help of sentiment analysis algorithms, which further provides an in-depth understanding of this particular problem domain. This project on sentiment analysis of movie reviews uses certain classifiers to govern the subjective opinion which further determines the polarity of the emotions expressed by the reviewer and its effect on the target audience. According to this research paper, among various classification techniques, the highest accuracy was provided by the Random Forest classification technique with an accuracy rate of 88.95%. This particular analysis gives scope for other domains of opinion mining like newspapers, articles, discussion forums, etc. in the future [5].

According to the study on Design Approach for accuracy in Movie Reviews using Sentiment Analysis by Rasika Wankhede and Prof. A N Thakare of Bapurao Deshmuk College of Engineering, the increased use of the internet and social media platforms for reviews, comments and feedbacks amongst users increases the scope of analyzing the reviews online by opinion mining to classify the emotions expressed by the reviewers. This approach further allows us to classify the polarity of subjective reviews as positive, negative and neutral. This particular research paper suggests more robust classifiers that can be used in place of conventional classifiers like SVM, Naïve Bayes and Maximum Entropy, which according to the paper have better results. This paper examines "Times of India" movie review database to perform sentiment analysis. They also have used Random Forest algorithm and achieved 90% accuracy. This paper makes it evident about how opinion mining helps the computational study of people's judgment and emotions towards a product or entity. The paper has used Input data, Pre-processing of texts using Tokenization, removal of Stop words, part of speech tagging, Text transformation algorithm, Feature extraction, and Feature reduction approach. By using these methodologies, the researchers have successfully classified the sentences in reviews online and done the feature impact analysis to understand the exact polarity results. This paper suggests how the NLP concept is being the best classification technique for achieving the highest accuracy in future works [10].

Charu Nanda, Mohit Dua and Garima Nanda focus on multilingual machine learning in their paper 'Sentiment Analysis of Movie Reviews in Hindi Language using Machine Learning'. This paper suggests the scope of Sentiment analysis in different languages. The paper uses Dataset reading and Pre-processing, Filtering based on positive and negative reviews, ignoring the neutral reviews, Classification using different algorithms, and performance evaluation using different metrics. This paper was able to classify reviews in Hindi into two classes, positive and negative and evaluating the same process using algorithms through different metrics, which help the user to distinguish positive and negative reviews in much less time. The scope suggests the use of neutral polarity and creating a base for comparing different algorithms for classification. All these suggest the scope of sentiment analysis and how it plays a vital role in resolving the common polarity issues of reviews on the internet [3].

3. Methodology

The work of sentiment analysis is carried out in the following process:

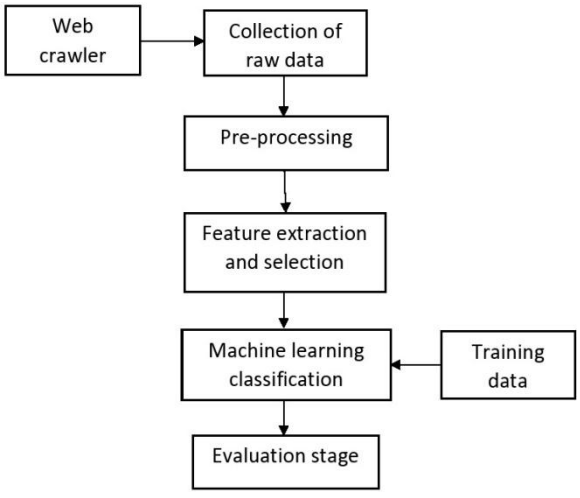

Figure 1: Architecture of the proposed work

A. Data reading and pre-processing

The input data is the set of reviews given by different people after watching a movie. Raw text can never be inserted to fit in machine learning. In order to clean text to fit in a machine learning model, we must clean the text first, which means the removal of punctuation and splitting into words etc.. This data is pre-processed in the following steps:

May-June 2020
ISSN: 0193-4120 Page No. 4427-4431

1. Splitting the text into words and converting them to lower case:

Each review contains a text which is made up of words. So, we split each text into words (as the words in the dictionary will not contain the entire text as a single word) and convert them to lower case as the upcoming steps will be easier to perform.

2. Stop words removal:

Words that have no meaning, such as "the", "is", "an" and do not provide any help in the analysis are known as stop words. Words such as I, me, myself, was, which, etc. which have no sentiments are also removed. Stop words are usually found in abundance thus they are usually removed before training the model. There is also a list of stop words which already exist in different languages. Stop words removal is a necessity for a better analysis.

3. Removing punctuation:

If there are any punctuation marks in the reviews used by the user, they will be removed as they do not add to the sentiment. Let us consider the removal of punctuations where characters like Full Stop (.), Question Mark (?), Double Inverted Commas (“ ”), Apostrophe (‘), Comma (,), Hyphen (-), dash [en dash (–) em dash (—)], Exclamation Mark (!), Colon (:), Semicolon (;), Parentheses (), Brackets [], Ellipsis (…), slash (/) are removed as they do not provide any meaning for analysis.

4. Stemming:

It is a process of reducing a word to its root word. For eg: Consider the words good, better and best, all the words have the same meaning, hence, the words better and best will be converted to good which is the root word.

5. Lemmatization:

This process is similar to stemming, but this process gives more accurate root words (as it identifies the meaning of the word in a sentence and also looks for the neighbouring sentences) than stemming and is also a better approach when compared to stemming as its efficiency is more.

B. Feature Extraction and selection:

Bag of words model is used for this purpose. This step gives the frequency of each word i.e., a unique word irrespective of its grammar and gives a set (bag) of words with its frequency (occurrence). The resultant is a {key: value} pair where the key is the unique word and value is its occurrence. The words after pre-processing will not contain any word that does not have a sentiment. Hence we select a few commonly used words (features) which are either positive or negative.

TF-IDF Vectorizer (frequency - inverse document frequency) gives the vector representation of each word. TF represents the number of times the word occurs and IDF represents how much information the word contains. It reduces the impact of the words which are less informative and have a high occurrence rate. These features are used to train the machine learning model.

C. Classification

Machine learning algorithms such as Decision Tree, Random forest, SVM etc. builds a mathematical model based on the training data. This process identifies to which class the feature belongs to. In our project, we identify the review given by the user is positive or negative. This is an important phase as we implement many algorithms to get the result.

Decision Tree is a type of supervised learning where the data is split based on a feature or a condition.

Random Forest is an ensemble learning method where multiple decision trees operate at the same time to predict the output.

Input to KNN (K nearest neighbours) consists of k nearest training values and the output depends on the vote of the nearest neighbours.

SVM (Support vector machine) is a type of supervised learning which builds a model to predict the output. This model consists of points in space thus separating one class from another.

Logistic Regression builds a model which contains the probability of a class. Each new prediction would be assigned a probability between 0-1 and their sum should be 1.

Naïve Bayes belongs to probabilistic classifiers and makes use of the Bayes theorem to predict the output. Input features are represented as vectors and the value of each feature is independent.

Adaptive Boost is used along with other machine learning algorithms (whose output is weak) to gain a better outcome.

D. Performance Evaluation

This is the last stage of sentiment analysis. Many measures can be used to measure the performance of the algorithms. The measures used in this paper are Precision, Accuracy, Recall, F1 score, Kappa, Average, Confusion matrix.

4. Results

This section presents assessment of various classifiers and performance measures used in this work.

TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 4427-4431

Table 1: Result of the proposed work

Algorithm used	Precision	Recall	Kappa Score	F1 Score	Average	Confusion Matrix
Decision Tree	71.34	70.90	41.86	71.12	**65.23**	[1581 647; 661 1611]
Random Forest	84.45	82.96	67.38	83.70	**80.43**	[1881 347; 387 1885]
KNN	74.62	79.97	52.27	77.20	**72.04**	[1610 618; 455 1817]
Naive Bayes	87.65	83.40	71.38	85.47	**82.72**	[1961 267; 377 1895]
Logistic Regression	88.01	89.83	77.37	88.91	**86.56**	[1950 278; 231 2041]
SVM	88.33	89.65	77.59	88.99	**86.67**	[1959 269; 235 2037]
Adaptive Boosting	80.71	83.05	62.83	81.86	**77.97**	[1777 451; 385 1887]

From Table 1 we can see that for SVM classifier yields the highest average of 86.67% except for Recall, for which Logistic Regression classifier yields the highest value of 89.83%.Therefore, SVM is considered the best among all the other classifiers according to the results.

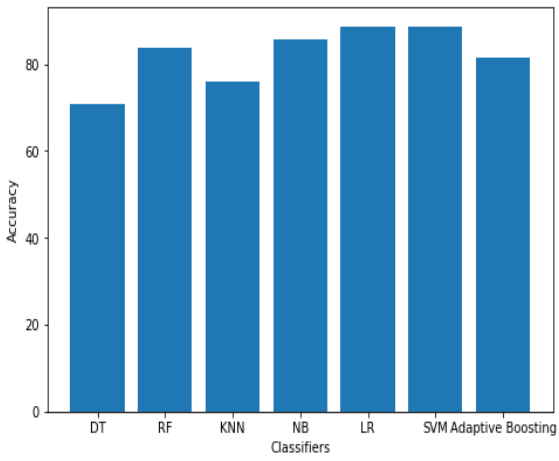

Figure 2: Result of the proposed work

From Fig. 2, we can see that SVM has achieved an accuracy rate of 88.8%, which is the highest among all the classifiers used in this work.

5. Conclusion

As per the study on the sentimental analysis, there have been various tasks that are done. Sentimental analysis has become very popular due to the increased number of users of social media and the internet. A lot of users go through a variety of movie reviews so that they can identify the quality of the movie and more than that they may have to spend a fair amount of time for the same. The relevance of this project is to classify sentences according to its sentiments by using different classification techniques.

In this project, we extracted the polarity of movie reviews through the classification method used in machine learning, which is a widely used method preferred over other methodologies. Through the classification method, the user has the advantage of saving time and get an accurate result on the reviews. Among the classification techniques, we found that SVM is more accurate because it gave us the highest accurate result of 88.8%.

In the future, the work can be further considered for studying the difference between various algorithms to find the suitable one for different tasks.

Reference

[1] Bhoir, Purtata, and Shilpa Kolte. "Sentiment Analysis of Movie Reviews Using Lexicon Approach." IEEE, Sentiment Analysis of Movie Reviews Using Lexicon Approach, 2015.

[2] Firmanto, Ari, and Riyanarto Sarno. "Prediction of Movie Sentiment Based on Reviews and Score on Rotten Tomatoes Using SentiWordnet." IEEE, *Prediction of Movie Sentiment Based on Reviews and Score on Rotten Tomatoes Using SentiWordnet*, 2018.

[3] Nanda, Charu, et al. "Sentiment Analysis of Movie Reviews in Hindi Language Using Machine Learning." IEEE, *Sentiment Analysis of Movie Reviews in Hindi Language Using Machine Learning*, 2018.

[4] Parkhe, Viraj, and Bhaskar Biswas. "Aspect Based Sentiment Analysis of Movie Reviews: Finding the Polarity Directing Aspects." IEEE, *Aspect Based Sentiment Analysis of Movie Reviews: Finding the Polarity Directing Aspects*, 2014.

[5] Sahu, Tirath Prasad, and Sanjeev Ahuja. "Sentiment Analysis of Movie Reviews: A Study on Feature Selection & Classification Algorithms." IEEE, Sentiment Analysis of Movie Reviews: A Study on Feature Selection & Classification Algorithms, 2016.

[6] Singh, V K, et al. "." IEEE, *Sentiment Analysis of Movie Reviews: A New Feature-Based Heuristic for Aspect-Level Sentiment Classification*, 2014.

May-June 2020
ISSN: 0193-4120 Page No. 4427-4431

[7] Singh, V K, et al. "Sentiment Analysis of Movie Reviews and Blog Posts." IEEE, *Sentiment Analysis of Movie Reviews and Blog Posts*, 2013.

[8] Timani, Heena, et al. "Predicting Success of a Movie from Youtube Trailer Comments Using Sentiment Analysis." IEEE, *Predicting Success of a Movie from Youtube Trailer Comments Using Sentiment Analysis*, 2019.

[9] Tripathi, Ankita, and Shrawan Kumar Trivedi. "Sentiment Analyis of Indian Movie Review with Various Feature Selection Techniques." IEEE, *Sentiment Analyis of Indian Movie Review with Various Feature Selection Techniques*, 2016.

[10] Wankhede, Rasika, and A N Thakare. "Design Approach for Accuracy in Movies Reviews Using Sentiment Analysis." IEEE, *Design Approach for Accuracy in Movies Reviews Using Sentiment Analysis*, 2017.

[11] Zhao, Kai, and Yaohong Jin. "A Hybrid Method for Sentiment Classification in Chinese Movie Reviews Based on Sentiment Labels." IEEE, *A Hybrid Method for Sentiment Classification in Chinese Movie Reviews Based on Sentiment Labels*, 2015.