iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** V **Month of publication:** May 2019

DOI: https://doi.org/10.22214/ijraset.2019.5200

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

Movie Review System Using Sentiment Analysis

Prachi Sanghvi[1], Disha Shah[2], Prof. Bharathi H. N.[3]
[1, 2, 3]*Computer Department, K. J Somaiya College of Engineering*

Abstract: ***The burgeoning presence of social media and open opportunities for posting personal views have flooded the internet with enormous volumes of opinions catering to all sorts of topics. This sheer volume of reviews has created some serious bottlenecks when it comes to gaining insight into any particular movie. Sentiment analysis is one domain of opinion mining that can solve this problem by taking reviews from multiple sources and condensing them into a single rating. In a movie review context, it is possible to use Sentiment Analysis to give the user a single rating for each movie that is a cumulative result of the reviews, comments that have been posted on various movie websites such as IMDB, TMDB, Rotten Tomatoes, Metacritic, etc. for that particular movie. Research done in this field has provided various techniques in machine learning that can be employed to achieve the classification of the sentiment in the review text as positive, negative or neutral. These methodologies involve feature extraction, creating a vocabulary of words for the system model and classifying them by applying appropriate algorithm like Naive Bayes and Support Vector Machine (SVM) to get more accurate results for a movie.***

Keywords: ***Sentiment analysis, TMDB, IMDB, Rotten Tomatoes, Metacritic, Naive Bayes, SVM, Movie Review.***

I. INTRODUCTION

The ever-increasing popularity of social media and other online platforms has enabled the public to make their opinions known worldwide. This has led to a massive amount of data being created on the various social networking sites in the form of tweets, comments, blogs, etc. However, as far as movie reviews are concerned, there are serious bottlenecks when the cost-conscious youth comes to making sense of these opinions. At the same time, the urgency to gain real-time updates has necessitated a condensed representation of this information. This project aims at obtaining feedback of the people from their comments on social media and then apply sentimental analysis. Sentiment analysis is a relatively new field in machine learning. It pertains to identifying the emotions or opinions that are expressed in the form text in some context. Here, it is applied to the comments with the help of machine learning algorithms like Naive Bayes and SVM. The accuracy of the classifier will be the deterministic factor in successfully predicting the sentiment after which the rating for a single movie can be easily calculated.

II. PROPOSED SYSTEM

The proposed system architecture has been shown in the following diagram. This is the core part of the system that focuses on the processing of the reviews to give a collective rating.

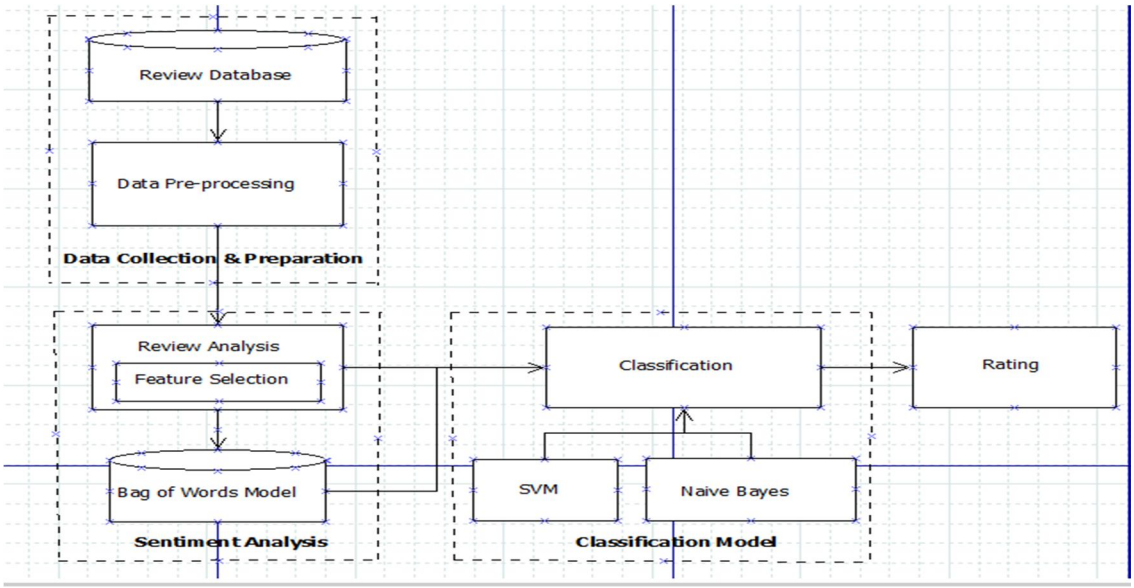


Fig 1. System architecture for movie recommendation system

Data preparation involves collecting and pre-processing user reviews for the subsequent analysis. A user review is likely to consist of unprocessed data. This means that the text can include stopwords, punctuations, blank spaces, etc. Sentiment analysis algorithms usually do not use the information other than the comments, that is the actual words that signify the sentiment or rating that the user/critic has conveyed. Feature generalization or metadata substitution is about generalizing features that may be overly specific. This task needs to performed so that a sentence such as "Padmaavat is a great movie", where "Padmaavat" is the name of the movie being reviewed, can be generalized in such a way that identifies "great movie" being responsible for conveying the positive sentiment of the review. Based on this principle, the review database that has already been segregated into positive and negative on the basis of their sentiment is processed and employed to construct a vocabulary of words. This vocabulary will contain the opinion words and their frequency of occurrence in the reviews of the selected dataset. The vocabulary that has been generated is used to train and test the classifier model. The classifiers being used are the Naive Bayes classifier and SVM, These are used to identify the opinion of unseen reviews of a particular movie and then give the output as a final rating for that movie.

III. IMPLEMENTATION

A. *Data Collection & Preparation*

The first step is to store the reviews that have been collected from authentic sources together form the review database that will be utilized in building the model. Now data needs to be prepared for feature selection. Pre-processing is applied to each review in the database. This involves removing tab spaces, newlines, dividing the sentences, removing numbers or digits, punctuation, removing stopwords (stopwords are the words that need to be filtered out before natural language processing), and converting all the letters to lowercase. Once this step is complete the output is stored in individual text files.

B. *Bag-of-Words Model*

The 'Bag-of-Words' model is a simple representation of the vocabulary that is being used to build the model. This is built form the review database. This bag of words is also used by the classifier to identify the features of input given and then classify it into the respective category. The bag of words is based on natural language processing (NLP) and information retrieval (IR). It can be thought of like a dictionary that stores distinct words along with a mapping of the words to their count; count is the frequency of occurrence of that particular word in the dataset that is used to build the model. The result is stored for later reference in a simple text file.

C. *Classification*

The vocabulary and review database together are used to train and test the classifier model. The partitioning of the main dataset for training and testing is done in the ratio of 60:40. Based on the words matched by the BoW model, the classifier acts as a binary classifier and will categorize the review/comment as positive or negative and assign a rating to each individual review. The average of the individual comments of a single movie is then calculated to arrive at a final rating for that particular movie.

1) *Naive Bayes Algorithm:* Naive Bayes is a simple but effective classification algorithm. The Naive Bayes algorithm is a widely used algorithm for document classification. The basic idea is to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories. The naive part of such a model is the assumption of word independence. The simplicity of this assumption makes the computation of Naive Bayes classifier far more efficient.
2) *Support Vector Machine (SVM):* Support vector machines (SVM), a discriminative classifier is considered the best text classification method. The support vector machine is a statistical classification method. Based on the structural risk minimization principle from the computational learning theory, SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set. Multiple variants of SVM have been developed in which Multiclass SVM is used for Sentiment Classification.
3) *Collecting Reviews for a Movie:* The movie for which a rating has to be obtained is decided. The reviews for this movie are searched on reliable online platforms like IMDB, Metacritic, Rotten Tomatoes, Tmdb. The reviews are scraped from these websites using an application called "Scrape Storm". These are extracted to a local database in the form of CSV files and then fed into the classifier to obtain a cumulative rating for the movie.

D. *Final Rating of a Movie*

The end result of any algorithm applied to the online reviews will be a rating. This rating is further used to arrive at a decision about whether the collective review for the movie has been positive, negative or neutral. This categorisation is done on the following basis:-

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

TABLE I
Categorization of Movies

Movie Rating (r)	Category
0 <= r < 2.5	Negative
r = 2.5	Neutral
2.5 < r <=5	Positive

IV. EXPERIMENTAL RESULTS

Two datasets of size 2,000 and 5,000 have been used for the training the algorithm each having subfolders labelled as positive and negative. Initially, these datasets were divided for training and testing in the ratio 80:20 in Naive Bayes classifier and obtained the accuracy of 74.99%. Later, after dividing training and testing dataset in the ratio of 60:40 accuracy was 78.59%. Hence, a standard ratio of 60:40 in training-testing is used for both the algorithms. The accuracies of both algorithms with the two datasets are given in the table below. Further, a review dataset of 25,000 reviews was also used for building the Naive Bayes model and the result accuracy of 84.74% was obtained which is the highest of all classifier models implemented in this project.
The accuracy of both the algorithm on both the dataset is as follows:-

TABLE II
Comparison of Accuracies Of The Algorithms

Dataset Size	Naive Bayes Accuracy (%)	SVM Accuracy (%)
2,000	78.59	84.37
5,000	79.19	80.85

TABLE III
Predictions for Sample Movies Of The Two Algorithms

Movie Name	Naive Bayes Rating	SVM Rating	Actual Rating
Chakde! India	4.2	2.8	4.1
Kalank	1.8	0.48	1.85
Dil Dhadakne Do	3.47	2.6	3.4

V. CONCLUSION & FUTURE WORK

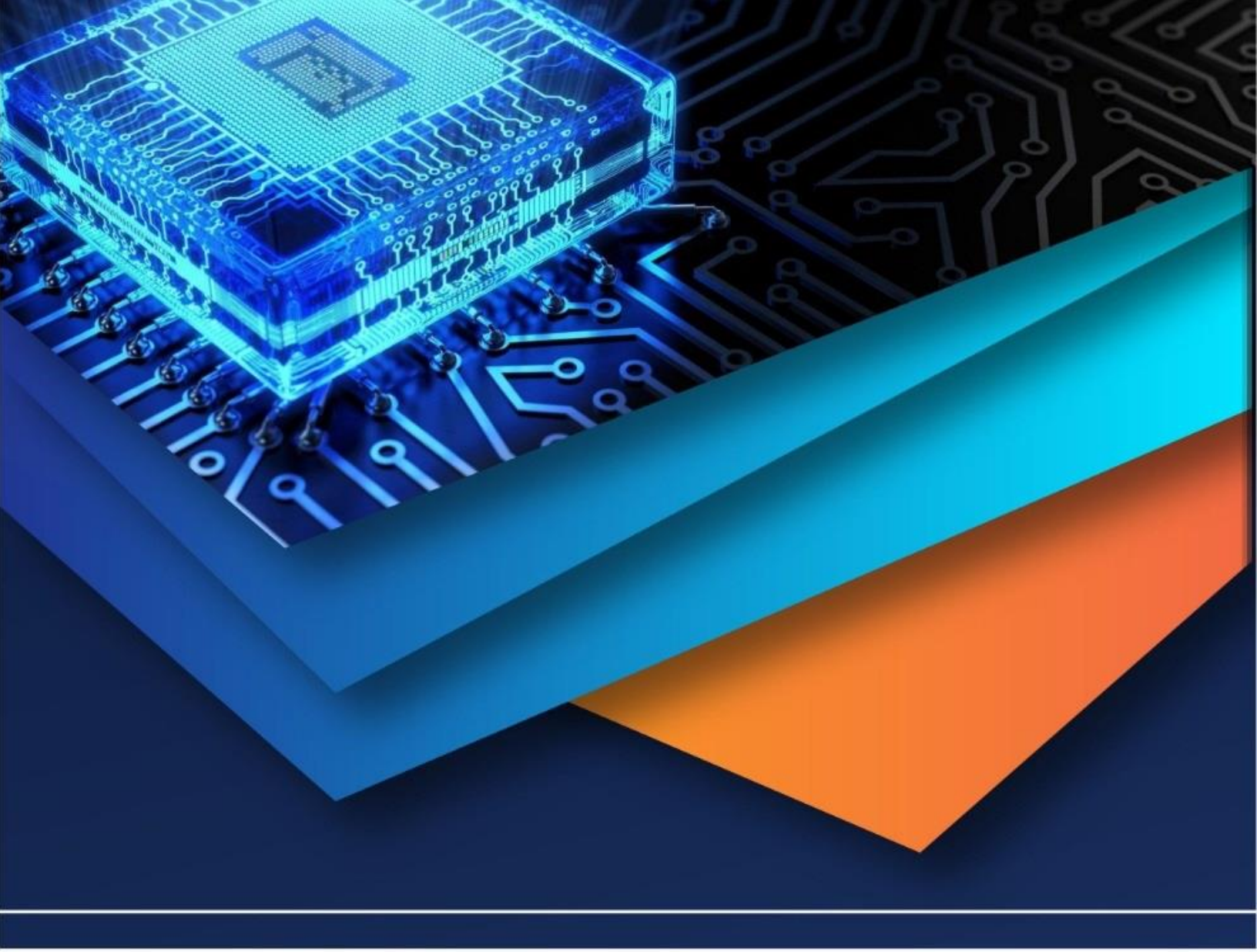
The paper mainly describes the implementation of two machine learning algorithms - Naive Bayes and SVM - that can be used to detect sentiments from the text. Naïve Bayes, being the more efficient algorithm provides the best accuracy and can be considered as a benchmark for all the other algorithms. Future scope of this project will be to increase the number of classes into which the sentiment can be classified. Here, only a binary classifier has been implemented but the same model can be extended to identify more than two sentiments if the required training and testing dataset can be found.

VI. ACKNOWLEDGEMENTS

The authors would like to thank their project mentor Prof. Bharathi H. N. for guiding them throughout the research of the project.

REFERENCES

[1] Tirath Prasad Sahu, Sanjeev Ahuja, "Sentiment analysis of movie reviews: A study on feature selection & classification algorithms." IEEE, 2016
[2] Mr B. Narendra, Mr K. Uday Sai, etc. "Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies." I.J. Intelligent Systems and Applications, 2016
[3] Jyotika Yadav, "A Survey on Sentiment Classification of Movie Reviews." IJEDR, 2014
[4] G. Hemantha Kumar, " Movie Recommendation based on Users' Tweets " Volume 141- No. 14, May 2016.
[5] G. Hemantha Kumar, "Movie Recommendation based on Users' Tweets " International Journal of Computer Applications, 2016
[6] Amrutha S Nair, Sreelakshmi K, " Movie Recommendation System Using Sentiment Analysis " International Journal for Trends in Engineering & Technology, 2017
[7] Palak Baid, Apoorva Gupta, Neelam Chaplot, "Sentiment Analysis of Movie Reviews Using Machine Learning Techniques" International Journal of Computer Applications, 2017
[8] G. Vinodhini, RM Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey" International Journal of Advanced Research in Computer Science and Software Engineering, 2012
[9] S. Sharma, D. Singh, "Study of Sentiment Classification Techniques" International Journal of Computer Sciences and Engineering, 2018
[10] Vibhor Singh, Priyansh Saxena, Siddharth Singh, S. Rajendran, "Opinion Mining and Analysis of Movie Reviews" Indian Journal of Science and Technology, 2017
[11] J Sai Teja, G Kiran Sai, M Druva Kumar, R.Manikandan, "Sentiment Analysis of Movie Reviews Using Machine Learning Algorithms - A Survey" International Journal of Pure and Applied Mathematics, 2018



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY