iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** V **Month of publication:** May 2021

DOI: https://doi.org/10.22214/ijraset.2021.34551

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: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

Movie Review System using NLP and SVM

Miss. Ghavate Rutuja[1], Miss. Morge Pooja[2], Miss. Borude Rushali[3], Mr. Pokharkar S T[4]
[1, 2, 3]*Computer Engineering Department, Shri Chhatrapati Shivaji College of Engineering, Rahuri Factory-413706, Savitribai Phule Pune University, Pune, Maharashtra.*

Abstract: ***Today we do a lot of things online from reading newspapers, watching entertainment and posting day to day thought about various things on social media. The postings on social media ranges from a news topic, personal views and review of movies we watched. These reviews represent true emotions after watching a movie and gives correct insight of the movie's watchable status. Thus, these kinds of reviews can be accessed and with the help of combination of latest technologies like machine learning and natural language processing we can get correct rating of a movie and help user decide whether the movie is good or bad. So, in this paper we propose a novel movie review system approach using ML and NLP together. To get reviews about a specific movie we are going to use IMDB and Movielens 100k dataset. First, we will accept a movie name as input and fetch movie reviews. Then stop words will be removed from the reviews using standard stop words database. Then the reviews will be analysed using OPEN-NLP tool and get lexicon analysis grammar from it i.e., whether a word is a verb, adjective, noun etc. Then we will keep only the necessary grammar. The analysed reviews will then be used to create a training dataset for ML algorithm. We are going to create a training dataset of our own. We propose to use SVM algorithm for predicting the movie in three categories bad, good and excellent. Thus, our system will be helpful to a user in getting correct movie review after which he can decide to watch a movie or not helping him to save money, time and mental stress by watching a bad movie.***
Keywords: IMDB, MovieLens, Review System, ML, NLP, SVM.

I. INTRODUCTION

This Today information from the internet has become an integral part of our daily life. We intend to do a lot of things on the internet from sharing information to managing entertainment needs. We use lot of social networking platforms like twitter, WhatsApp, IMDb etc. to share information about our daily experiences and events we come across. But the sheer volume of information it becomes very complicated for a human brain to keep track of. Due to blast of information, we are not able to understand the fake and correct information. To analyses good and bad information we have to give lot of time to do so. Thus, if an information is not correctly filtered it will cost us both money and time. Basically, two types of filters collaborative filter and content filter can be used to filter data on the internet. In collaborative filtering data filtering can be done using views of various users together. In content-based filtering data is filtered using the content written on the internet about a specific topic. Thus, Fig. 1 shows the filtering process.

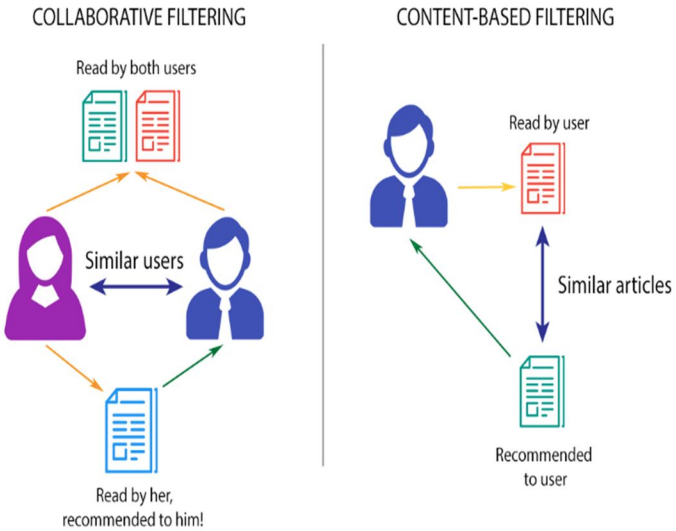


Fig. 1 Existing filtering techniques.

Thus, by studying the filtering processes we come to a understanding that conent based filtering has more potential as lot of data is available in content section. Thus content based data can be found in abundance. We also need to combine Natural Language Processing (NLP) and Machine Learning (ML) technologies together to implement proper content based filtering. Thus using NLP and ML the filtering process should filter the data in three categires bad, good and excellent for better understading. So to design a better movie review system we need to study various ideas and reasearch put forth by other authors and try to understand the techniuqes used by them and how to improve it in our proposed system. So, in other words the main objective of the paper is to:

A. Give correct insight on the watchable status of a movie.
B. To help a user to access large amount of movie reviews without human interaction.
C. Study and implement a good movie review system using NLP and ML.
D. To study previous techniques mentioned by other authors to get an improved movie review system.
E. Proper training dataset should be created for each movie.
F. Evaluate and analyze the new movie review framework and understand its strength and drawbacks.

Thus, the rest of the paper is structured as follows:
1) Section II. explains literature survey which studies previous techniques used by other authors.
2) Section III. explains problem definition with goals of the new framework used to create a movie review system.
3) Section IV. explains the methodology i.e., mathematical model and algorithms to be used by the new movie review system.
4) Section V. explains proposed system with detailed working and system architecture of the system.

II. LITERATURE SURVEY

This section describes the various techniques and the study associated with them that can be used in designing a new movie review system. It helps in understanding various ideas put forward by various technical papers published by various authors and how they put forth a more accurate and concrete techniques. Some of the ideas with technique and drawbacks are mentioned below:

In 2020 Sandesh Tripathi et al. [1] presented the paper which focusses mainly on analysing movie review sentiments using Naïve Bayes and Linear Regression. This technique is quite good and covers all the things needed for a successful movie review system. But main drawback of this system is that it concentrates on IMDB dataset for all movies designed commonly for training and does not design a new training dataset for each movie specifically.

In 2020 Nimish Kapoor et al. [2] presented the paper which focusses mainly on analysing movie recommendation system using NLP and SVM. This technique is quite good and covers all the things needed for a successful movie review system. But main drawback of this system is that it concentrates on TMDB dataset for all movies designed commonly for training and does not design a new training dataset for each movie specifically.

In 2020 Oumaima Hourrane et al. [3] presented the paper which focusses sentiment classification on movie reviews using Linear Regression and SVM. This technique is quite good and covers all the things needed for a successful movie review system. But main drawback of this system is that it concentrates on twitter and IMDB dataset for all movies designed commonly for training and does not design a new training dataset for each movie specifically.

In 2019 Mara-Renata Petrusel et al. [4] presented the paper which focusses mainly on restaurant recommendation system using Naïve Bayes and SVM. This technique is quite good and can be used to create a movie review system. But main drawback of this system is that it concentrates on Yelp Restaurants Reviews dataset for all recommendation and does not design a training dataset of its own.

In 2020 Sai Chandra Rachiraju et al. [5] presented the paper which focusses mainly on feature extraction and classification of movie reviews using Linear SVM and Random Forest. This technique is quite good and covers all the things needed for a successful movie review system. But main drawback of this system is that it concentrates on IMDB dataset for all movies designed commonly for training and does not design a new training dataset for each movie specifically.

III. PROBLEM STATEMENT

All Movie review and recommendation system is a hot topic. A large amount of movie review data can be found on the internet on movie database sites IMDB, Twitter, MovieLens etc. This database due to large in nature is hard to analyse by a human interaction and thus prone to errors resulting in wrong analysis. Many production houses give wrong reviews of movies by paying to a reviewer.

Thus, the reviewer gives false reviews and increase rating of an undeserving movie. If movie reviews are wrongfully analysed a user can fall prey to wrongful viewing of a movie which can result in to loss of time and a lot of mental torment due to undeserved movie watching. So, there is need of a system where a user can analyse movie reviews correctly and save mental tormenting, money and time. So main goals of our system can be stated as:

A. To design and execute new movie review system framework effectively.
B. To make effective use of more than one movie databases.
C. To make use of movie reviews from MovieLens and IMDB databases.
D. To analyse the movie data using NLP and ML technologies together.
E. To use standard stop words removal database.
F. To make effective use of OPEN-NLP tool to implement lexicon analysis on reviews.
G. To predict movie's watchable status in three categories such as bad, good and excellent.

IV. METHODOLOGY

This section will explain the algorithm and mathematical conditions to be used for designing an automatic and accurate movie review system. Thus, they can be explained as follows:

A. Mathematical Model

Our movie review system technique can be explained in two sets with probability, success and failure conditions.

1) Preprocessing Module

Set (P)= {P0, P1, P2, P3, P4, P5}

$P0 \in\in P$ = Enter movie keyword.

$P1 \in\in P$ = Fetch movie reviews from IMDB dataset.

$P2 \in\in P$ = Remove stop words using standard stop words array.

$P3 \in\in P$ = Perform lexicon analysis using OPEN-NLP.

$P4 \in\in P$ = Extract features and keep valid words.

$P5 \in\in P$ = View results.

2) SVM Module

Set (S)= {S0, S1, S2, S3, D5}

$S0 \in\in S$ = Create training dataset using extracted features words.

$S1 \in\in S$ = Pass a test movie reviews and create testing dataset.

$S2 \in\in S$ = Perform classification using training and testing dataset with SVM algorithm.

$S3 \in\in S$ = Predict review in three classes good, bad and excellent.

$P5 \in\in S$ = View results.

So, by studying the sets we come to notice that elements are common in both modules and used in coordination in both sets so they be placed as

$$x \in\in P \cap S \text{ if } x \in\in P \text{ and } x \in\in S$$

Thus, the probability of intersection of elements in both modules can be given as

$$P(P \cap S) = P(P)+P(S)$$

So, intersection of common elements can be shown as

$$P \cap S = \{P5\}$$

The conditional probability of both modules using the same element can be shown as

$$P(P \mid S) = \frac{P(P \cap S)}{P(S)}$$

Thus, we conclude that our movie review framework's success and failure will depend upon the internet to get movie reviews, i.e., if the internet is not good or not present, we will not be able to download movie reviews and the project will not work, thus this is a case of failure, so our framework supports NP-Hard and not NP-Complete.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

B. Algorithms Used

The ML algorithm SVM will be used for successful implementation of the movie review system.

1) *SVM:* This algorithm will be used to classify movie reviews in three classes bad, good and excellent. It is a machine learning algorithm called as support vector machine. It is a supervised learning model. It is used with associated learning algorithms that analyze data used for classification and regression analysis. If two categories are given for classification using SVM then an SVM training algorithm builds a model with two planes and places each category in one plane or the other, making it a non-probabilistic binary linear classifier. An SVM model can be representation of points in space mapped in such a way that the examples of each separate categories are placed in a separate plane by a clear gap that is wide as possible. It needs a training and testing dataset for classification. It needs to be trained properly for better results.

V. PROPOSED SYSTEM

This section explains the propose system with the help of system architecture diagram and working as shown in Fig.2. The working modules are mainly divided in three parts as follows:

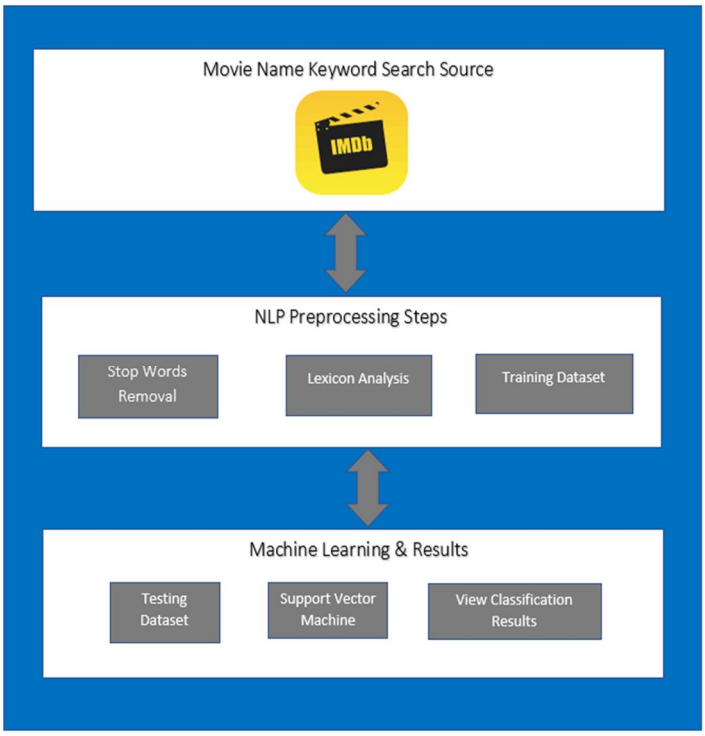


Fig.2. System Architecture Diagram.

A. Movie Review Data collection Using IMDB

This module will be proposed to collect data needed for the movie review system to work on. The data will be collected using IMDB API. The data from the IMDB database will be collected using a movie keyword and stored as txt files on local drives for further use.

B. NLP Based Pre-processing

This module will pre-process data to implement ML algorithm SVM. First the data collected in txt files will be called and stop words will be removed from the data using standard stop words database. The stop words removed data will then be passed to OPEN-NL tool to apply lexicon analysis. The Lexicon analysis will tell if a word is verb adjective, noun etc. Only the relevant grammar type will be kept. The data will then be used to create a training dataset which will be used to train SVM algorithm.

C. Machine Learning

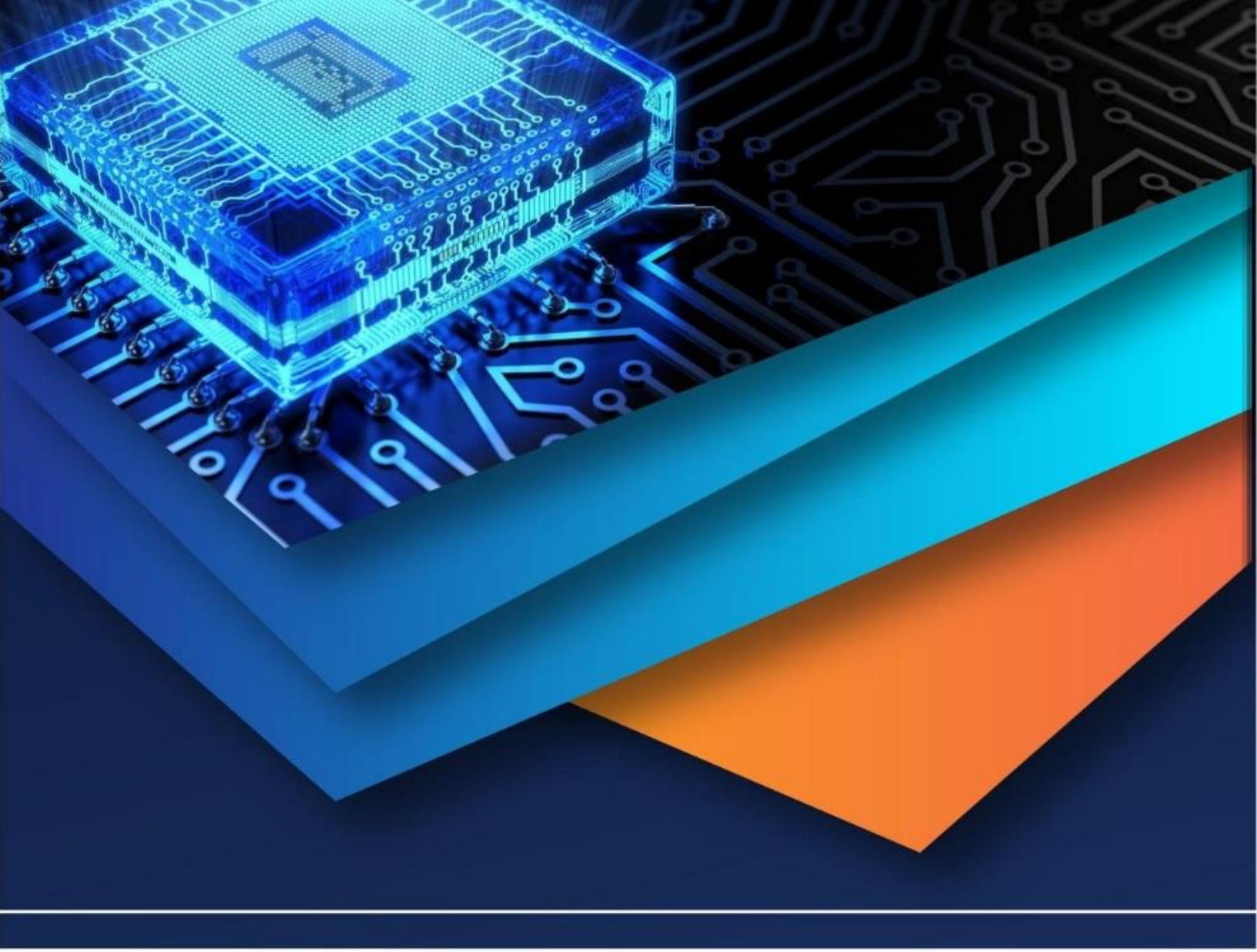
This module will first create a testing dataset which will be created using a test review. The SVM algorithm will be trained using a training dataset. After training the SVM algorithm will be applied on testing dataset which will return prediction results in three categories like good, bad and excellent. Thus, it shows the review category in front of it.

VI. CONCLUSIONS

The Thus, in this paper we conclude that we have created a successful movie review system using NLP and SVM together. For successful proposal of our system we took in to consideration the works done by [1][2][3][4][5] authors and tried to incorporate in our new system. We are using IMDB as our movie review database and download reviews from it using IMDB API. To increase prediction accuracy for a movie review are creating training dataset of our own for each movie. We have used standard stop words database to remove stop words. We have used OPEN-NLP tool to apply lexicon analysis and keep only relevant words. We have used ML supervised learning algorithm SVM which can give prediction with high accuracy in spite of small training dataset. Thus, we conclude that our new movie review framework will help save a user's money and time by not watching a bad movie.

REFERENCES

[1] Sandesh Tripathi, Ritu Mehrotra, Vidushi Bansal and Shweta Upadhyay, "Analyzing Sentiment using IMDb Dataset" in IEEE-2020.
[2] Nimish Kapoor, Saurav Vishal and Krishnaveni K S, "Movie Recommendation System Using NLP Tools" in IEEE-2020.
[3] Oumaima Hourrane, Nouhaila Idrissi and El Habib Benlahmar, "Sentiment Classification on Movie Reviews and Twitter: An Experimental Study of Supervised Learning Models" in IEEE-2020.
[4] Mara-Renata Petrusel and Sergiu-George Limboi, "A Restaurants Recommendation System: Improving Rating Predictions using Sentiment Analysis" in IEEE-2019.
[5] Sai Chandra Rachiraju and Madamala Revanth, "Feature Extraction and Classification of Movie Reviews using Advanced Machine Learning Models" in IEEE-2020.
[6] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, - Enhancing deep learning sentiment analysis with ensemble techniques in social applications, Expert Syst. Appl., vol. 77, pp. 236–246, Jul. 2017.
[7] E. Aslanian, M. Radmanesh, and M. Jalili, - Hybrid recommender systems based on content feature relationship, IEEE Trans. Ind. Informant., early access, Nov. 21, 2016, doi: 10.1109/TII.2016.2631138.
[8] Yasen, Mais, and Sara Tedmori. -Movies Reviews sentiment analysis and classification, IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2019.
[9] Mishne, Gilad and Natalie Glance, (2006), -Predicting Movie Sales from Blogger Sentiment, AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs.
[10] Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath, (2016), -Classification of sentiment reviews using n-gram machine learning approach, Expert Systems with Applications, Vol. 57, PP. 117-126.
[11] Vivek Narayanan, Ishan Arora, Arjun Bhatia, (2013), -Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model, Intelligent Data Engineering and Automated Learning – IDEAL, Springer, Vol. 8206.
[12] Hossin, Mohammad, and M. N. Sulaiman. -A review on evaluation metrics for data classification evaluations, International Journal of Data Mining & Knowledge Management Process 5.2 (2015): 1.
[13] Kerstin Denecke, (2008), -Using SentiWordNet for multilingual sentiment analysis, IEEE 24th International Conference on Data Engineering Workshop, Cancun, PP 507-512



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY