

# Interpreting Movie Reviews from Public Comments

**Abstract**—Analysing the movie reviews helps to understand whether a movie is loved by the public or not which determines the success rate of the movie. The movie reviews contain the sentiments of the public that they express to convey their opinion regarding the movie. In this study, to analyze these sentiments five different machine learning methods are used to classify reviews as either positive or negative. Our model is tested using a real-world dataset and the five methods analyze and compare the model. After performing the five algorithms on our model it is seen that SVM and Naive Bayes outperform the other three models. On the other hand, Decision Tree Machine Learning gives an unpleasant or lowest outcome on the dataset.

## I. INTRODUCTION

Movie reviews are essential for understanding the impact of a film on its audience. On its website, IMDB gives rating methods. However, not everyone rates a film. We may learn a lot more about the movie's reception by reading the comments sections on various social media websites. Comments and postings on a movie provide us with a more in-depth understanding of the film's various components.

The purpose of this paper is to analyze the reviews of the audience on the IMDB movies and classify them into different sentiments. The sentiment from data may be used to determine the public opinion of a movie. Analyzing audience perceptions will help to determine the kinds of movies they prefer. Moreover, we are using IMDB comments because IMDB has a diverse source of opinions and sentiments that can be from the most recent movies the audience watched to a movie they have previously seen. The audience's review comments include a wide range of topics, including politics, religion, and even the individual's mental state. This study focuses on sentiment analysis of IMDB reviews. Machine learning techniques are employed to extract sentiment from data.

Sentiment analysis is an excellent method for determining public opinion. In this method, we process the dataset to remove irrelevant information. For example, punctuation marks do not tell us much about the type of review. In the first phase we did pre-processing. Where we cleaned the punctuation marks. Furthermore, 'stopwords' such as 'the,' 'and,' 'is,' and so on are unnecessary words that can lead the machine learning algorithm misguided. We must additionally check for null values, delete blank spaces, and eliminate new lines. Therefore, these superfluous parts were also omitted during the pre-processing step. Thus, only the data relevant to the study is acquired for the next phases. Stemming is a crucial aspect of data preprocessing. We stem words by converting them to their base words. More features are retrieved and added to the feature vector in the following step. To Train the model we utilized logistic regression, support vector classification, naive

bayes, random forest and decision tree. After training, when the model is ready, finally test reviews are fed into the model, and finally, we have reviews that have been categorized

Machines are incapable of distinguishing between reviews that are positive, negative, or neutral. They cannot also be trained to do so. Understanding a person's mental state from their reviews is therefore an impossible task for a machine to perform. However, it is simpler to determine the sentiment of the review when using various machine learning algorithms. Hence, these lead us to our objectives which are:

- Training machine so that it can overcome the barrier of grammar conventions, dialects and mistyped words.
- Extracting relevant data from the reviews.
- Researching the most used words in the reviews.
- Generating scores using classification algorithms.

## II. RELATED WORKS

Abdül MERAL, a data scientist at LC WAIKIKI in Istanbul, Turkey, did the finest job with this dataset. To anticipate movie reviews, he utilized deep learning and a neural network model. His model has an accuracy of 89.999998%. Diksha Bhati, a Delhi University student, did another work with this dataset. Her model analyzed the reviews with cnn-pytorch. This model has an accuracy of 88%.

## III. DATASET

[5] [2] The dataset has 40000 Movie reviews where most of the reviews are longer than 200 words. A positive review is labeled as 0, and a negative review is labeled as 1. Among them, 32000 reviews are used for training the machine learning models and 8000 reviews are used for testing the models.

## IV. METHODOLOGY

Now, we will be discussing the whole process of our work on interpreting movie reviews from public comments. Here, we have done dataset balancing, preprocessing (null checking, stemming etc.), vectorization, dataset splitting and finally we trained the ML models using the dataset.

The steps followed are shown in the "Fig. 1".

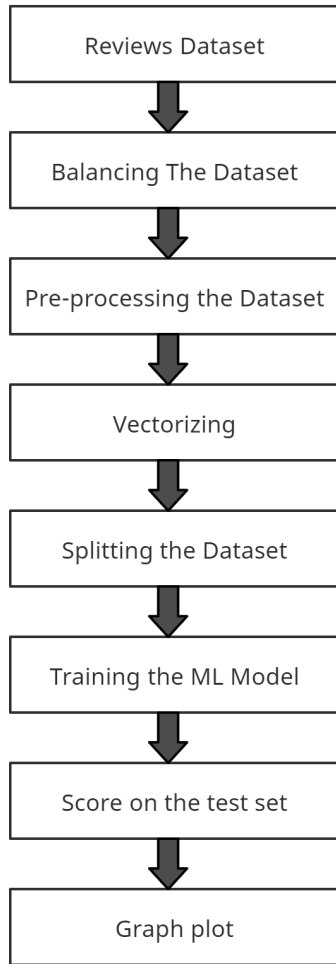


Fig. 1. Steps Followed

#### A. Balancing the Dataset

We checked if the dataset is balanced or not. For example, if we have 2000 positive reviews and 14000 negative reviews for training, our algorithm will not be able to learn the positive reviews properly. In our dataset we had 20019 positive reviews and 19981 negative reviews. That means 50.1% positive reviews and 49.9% negative reviews. That means our dataset is already balanced.

#### B. Preprocessing Data

It is an important part to teach the machine learning algorithms. Otherwise the algorithm maybe misguided. To preprocess the data, we took the following steps.

- Null Checking - Null values do not have any labels.
- Removed link - Links don't tell us about the sentiment of the review.
- Removed line breaks - Removing line breaks makes the dictionary of words smaller and makes it time efficient and more accurate for the machine to learn.
- Removed extra spaces

- Removed punctuation - punctuation marks don't tell us much about the type of review.
- Removed stopwords - 'Stopwords' such as 'the,' 'and,' 'is,' and so on are unnecessary words that can lead the machine learning algorithm misguided.
- Stemming - Stemming is a crucial aspect of data pre-processing. We stem words by converting them to their base words. For example, after stemming, the word 'Connected' becomes 'Connect'. Different variants of the same words that mean the same thing can be found in the English language. However, an ML system will recognize the words 'Connect' and 'Connected' differently, and the dictionary size will be enormous as a result. That is why we transform all words to their root form, reducing the dictionary size and improving the ML algorithm's performance.
- Lowercased the reviews - Machines detect 'LOVE' and 'love' differently as they are spelled differently. If we convert all the words into lower cases, then it is easier for the machine to learn.

#### C. Vectorization

Vectorization is a method for converting text input into real-world numerical data. By transforming the text into numerical vectors, we were able to extract several distinguishing characteristics from the text for training purposes. We have used TF-IDF vectorization on the dataset where we multiplied two different matrices.

#### D. Splitting the Dataset

The dataset was divided before training. We kept 80% of the dataset for training, and the remaining 20% for testing.

#### E. Training the ML Models

##### • LOGISTIC REGRESSION

[3] Logistic regression is a classification model algorithm of supervised machine learning. Firstly, supervised learning means labeled data and the purpose is to predict the value of a model and classification model is an algorithm of supervised learning which predicts the outcome of a model if something will happen or not. Logistic regression is used to calculate the probability of an event occurring [12]. The sigmoid function of  $f(x)$  :  $p(x) = 1/(1+\exp(f(x)))$  is used for logistic regression. The formula of logistic regression is : [8]

$$h_{\theta}(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \quad (1)$$

The result of this formula is then used as a probability value with the help of the given sigmoid function.

##### • SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a binary classification model and uses kernel function. Here, we plot the classes in an n -Dimensional space. The main intention of SVM model is to calculate a hyperplane which will differentiate the classes plotted [13].

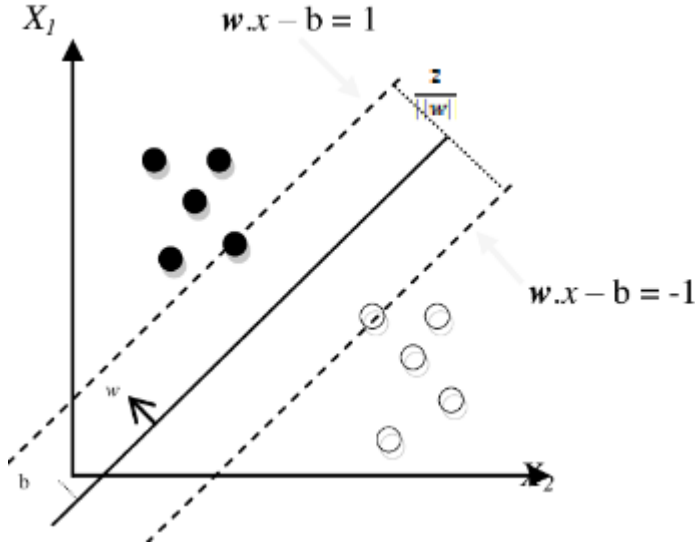


Fig. 2. SVM [9]

The cost function of SVM is showed below: [7]

$$J(\theta) = \frac{1}{2} \sum_{j=1}^n \theta_j^2 \quad (2)$$

$$\theta^T x^{(i)} \geq 1, \quad y^{(i)} = 1 \quad (3)$$

$$\theta^T x^{(i)} \leq -1, \quad y^{(i)} = 0 \quad (4)$$

If we minimize the value of theta, we can increase the accuracy of the algorithm. The cost is calculated by the kernel functions.

#### • NAIVE BAYES

[3] In this algorithm, we use the Bayes rule, which is like the following.

$$P(X | Y) = (P(Y)P(Y | X)) \div P(X)$$

Here, we find the ratio of the number of the positive reviews and negative reviews. Then we find the probability of the word given the review is positive or negative. Naive Bayes is better suited to category variables than numerical variables. When we need to predict data, we use it [14].

#### • RANDOM FOREST

It is a supervised machine learning algorithm which is used for classification and regression problems. This algorithm is a combination of multiple models, not a single model [11]. The random forest algorithm involves the following steps:

Step 1: We choose n number of random records from k number of records

Step 2: An individual decision tree is built.

Step 3: Every decision tree will give us outputs.

Step 4: The final output is based on majority voting or averaging.

Random forest generates a large number of decision trees while training. To classify, we choose the output of the

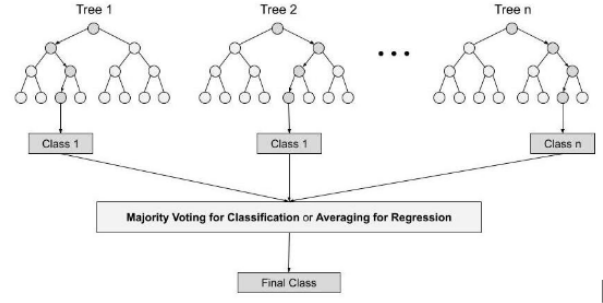


Fig. 3. Random Forest Algorithm [10]

majority of the trees. For regression tasks, we choose average prediction.

#### • DECISION TREE

Decision tree is also a method that is used to solve classification problems. However, in contrast to other supervised machine learning techniques, decision trees also solve the problems regarding regression. It is a classifier with a tree structure. This type of supervised machine learning is used for developing training models, which can later be used to find out the class or value of the target variable through learning simple decision rules derived from training data.

Here, the decision tree classifier is used to predict a class label from the dataset. To do that, we begin at the tree's root. Then came the comparison part. We differentiated between the roots and the values that are in the records. We traveled to the next nodes depending on which branch node correlates to that value [11].

## V. RESULT

From the result of our experiment it is seen that Logistic Regression (LR) gives 88% accuracy, SVM gives 89% accuracy, Naive Bayes also gives 89% accuracy, Random Forest (RF) gives 86% accuracy and lastly Decision Tree gives 72% of accuracy. So from the results of our experiment, we can say that The SVM and Naive Bayes are the most effective algorithms for sentiment analysis of IMDB movie reviews as both give the highest accuracy possible. LR and RF also give a good amount of accuracy. However, the accuracy level of the Decision Tree is not satisfactory. So we can conclude that SVM, Naive Bayes and Logistic Regression are the effective algorithms for Sentiment Analysis of IMDB Movie Review.

TABLE I  
ACCURACY OF THE ML MODELS

Algorithm	Accuracy	f1-score
Logistic Regression	0.88	0.88
Support Vector	0.89	0.89
Naive Bayes	0.89	0.88
Random Forest	0.86	0.86
Decision Tree	0.72	0.72

## VI. CONCLUSION

This article demonstrates our work on Sentiment Analysis of IMDB Movie Reviews. Our findings indicate that machine learning is the best approach for sentiment analysis. Our studies' accuracy results reveal that SVM, Naive Bayes, and Logistic Regression are the most effective machine learning algorithms for predicting the sentiments of IMDB movie reviews.

We discovered which machine learning models are the most effective in predicting movie reviews as a result of this experiment. We can understand the impact of a movie on the audience by anticipating the type of movie reviews based on public debates.

## REFERENCES

- [1] Sharma, P. "An Introduction to Stemming in Natural Language Processing." *Analytics Vidhya*, November 2021. [shorturl.at/DFJPZ](https://shorturl.at/DFJPZ).
- [2] "IMDB dataset (Sentiment analysis)" *Kaggle* <https://www.kaggle.com/datasets/columbine/imdb-dataset-sentiment-analysis-in-csv-format?resource=download&select=Valid.csv>.
- [3] Logistic Regression and Naive Bayes. "Sentiment Analysis using Logistic Regression and Naive Bayes." , 28 Nov. 2020, <https://towardsdatascience.com/sentiment-analysis-using-logistic-regression-and-naive-bayes-16b806eb4c4b>.
- [4] Contributor, TechTarget. "What Is Support Vector Machine (SVM)? - Definition from WhatIs.com." *WhatIs.com*, TechTarget, 29 Nov. 2017, <https://www.techtarget.com/whatis/definition/support-vector-machine-SVM>.
- [5] Goyal, A. and Parulekar, A., 2022. Sentiment Analysis for Movie Reviews. [online] *Cseweb.ucsd.edu*. Available at: <https://cseweb.ucsd.edu/classes/wi15/cse255-a/reports/fa15/003.pdf> [Accessed 24 August 2022].
- [6] Analytics Vidhya. 2022. Random Forest — Introduction to Random Forest Algorithm. [online] Available at: <https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/> [Accessed 5 September 2022].
- [7] freeCodeCamp.org. 2022. SVM Machine Learning Algorithm Explained. [online] Available at: <https://www.freecodecamp.org/news/support-vector-machines/>; :text=The
- [8] Medium. 2022. Twitter sentiment analysis using Logistic Regression. [online] Available at: <https://medium.com/nerd-for-tech/twitter-sentiment-analysis-using-logistic-regression-ff9944982c67> [Accessed 20 August 2022].
- [9] Medium. 2022. Sentiment Analysis using SVM. [online] Available at: <https://medium.com/@vasista/sentiment-analysis-using-svm-338d418e3ff1> [Accessed 20 August 2022].
- [10] The Science of Machine Learning. 2022. Random Forest — The Science of Machine Learning. [online] Available at: <https://www.ml-science.com/random-forest> [Accessed 17 August 2022].
- [11] @article{article, author = Ali, Jihad and Khan, Rehanullah and Ahmad, Nasir and Maqsood, Imran, year = 2012, month = 09, pages = , title = Random Forests and Decision Trees, volume = 9, journal = International Journal of Computer Science Issues(IJCSI)}
- [12] @article{article, author = Peng, Joanne and Lee, Kuk and Ingersoll, Gary, year = 2002, month = 09, pages = 3-14, title = An Introduction to Logistic Regression Analysis and Reporting, volume = 96, journal = Journal of Educational Research - J EDUC RES, doi = 10.1080/00220670209598786}
- [13] TY - BOOK AU - Evgeniou, Theodoros AU - Pontil, Massimiliano PY - 2001/01/01 SP - 249 EP - 257 T1 - Support Vector Machines: Theory and Applications VL - 2049 DO - 10.1007/3-540-44673-7\_12 ER - @article{article, author = Rish, Irina, year = 2001, month = 01, pages = , title = An Empirical Study of the Naïve Bayes Classifier, volume = 3, journal =