Sentiment Analysis of IMDB Reviews

Md Raihanul Islam Bhuiyan
Undergraduate, Computer Science and Engineering
Brac University
Dhaka, Bangladesh
raihanul.islam.bhuiyan@g.bracu.ac.bd

Mahin Shahriar Efaz
Undergraduate, Computer Science and Engineering
Brac University
Dhaka, Bangladesh
mahin.shahriar.efaz@g.bracu.ac.bd

Aditi Saha Ria
Undergraduate, Computer Science and Engineering
Brac University
Dhaka, Bangladesh
aditi.saha.ria@g.bracu.ac.bd

***Abstract*—We use sentiment analysis to evaluate whether a Natural Language is Positive, Negative, or Neutral. For example, on social media, we see a lot of remarks that are either motivating or abusive. We can determine the nature of text using NLP algorithms and sentiment analysis.**

In our research, we gathered a dataset of IMDB comment sections. It is an internet database where we can view information about movies, short films, and television shows. We will use sentiment analysis to determine whether the movie review is good or negative. We'll utilize various machine learning algorithms to detect movie reviews and see which one gives us the best results.

A. *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 don't 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

I. Objectives of The Study

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. STEPS FOR SENTIMENTAL ANALYSIS

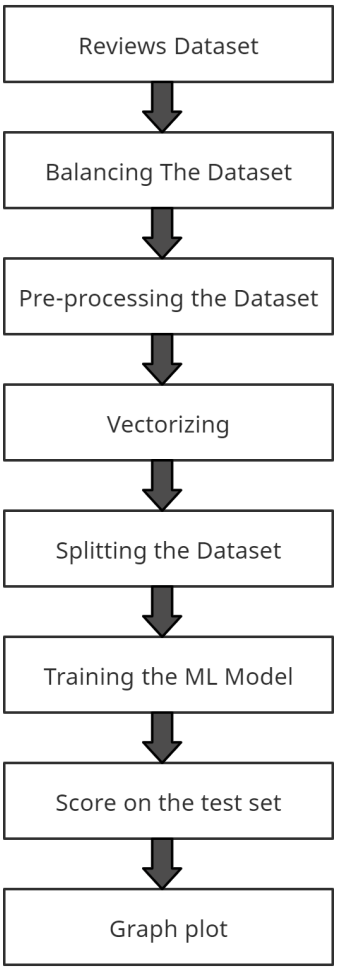

Fig. 1. Steps Followed

A. *Dataset*

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.

B. *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.

C. *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 preprocessing. 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.

D. *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 matrics.

E. *Splitting the Dataset*

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

F. *Training the dataset with ML models:*

- **Logistic regression:** 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. The sigmoid function of f(x) : p(x) = 1/(1+exp(f(x)) is used for logistic regression. The formula of logistic regression is :

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

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.

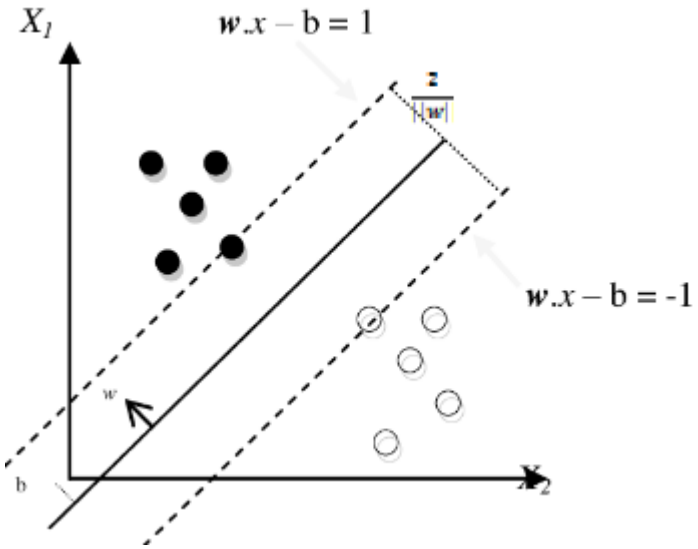

Fig. 2. SVM

The cost function of SVM is showed below:

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

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

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

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:** In this algorithm, we use the Bayes rule, which is like the following.

$$\mathrm{P(X} \mid Y) = (P(Y)P(Y \mid X)) \div P(X)$$

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.

- **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. The random forest algorithm involves the following steps:

Step 1: In Random forest, n random records are chosen at random from a data collection of k records.
Step 2: For each sample, an individual decision tree is built.
Step 3: Each decision tree will produce a result.
Step 4: For classification and regression, the final output is based on majority voting or averaging.

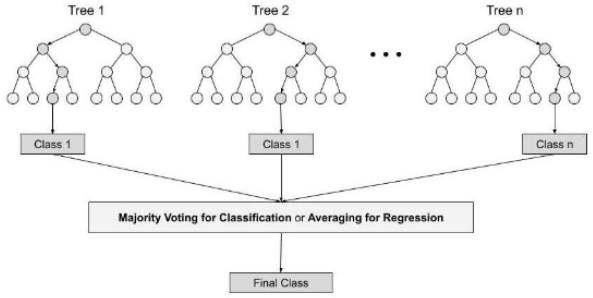

Fig. 3. Example of a figure caption.

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

- **DECISION TREE**

Decision tree is one of the supervised learning algorithms. The decision tree approach, in contrast to other supervised learning algorithms, may also be utilized to solve classification and regression. It is a classifier with a tree structure.Using a Decision Tree, the objective is to develop a training model that can be used to determine the class or value of the target variable by learning straightforward decision rules inferred 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, we compare the root attribute's values with that of the attribute on the record. We follow the branch that corresponds to that value and go on to the next node based on the comparison.

G. 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% accuray, 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.

ML Algorithms Used	
	Accuracy
Logistic Regression	0.88
Support Vector Machine	0.89
Naive Bayes	0.89
Random Forest	0.86
Decision Tree	0.72

*H. **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.

[2] "IMDB dataset (Sentiment analysis) " *Kaggle* https://www.kaggle.com/datasets/columbine/imdb-dataset-sentiment-analysis-in-csv-format?resource=downloadselect=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] Ankit Goyal, Amey Parulekar,. "Sentiment Analysis for Movie Reviews" *Sentiment Analysis for Movie Reviews* https://cseweb.ucsd.edu/classes/wi15/cse255-a/reports/fa15/003.pdf.

[6] Random Forest. "Understanding Random Forest" , 21 Apr. 2021, https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/.