International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

SENTIMENTAL ANALYSIS FOR IMDB USER REVIEWS

1. Hitesh Salunkhe 2. Ashutosh 3. Nikhil Vilhekar 4. Pramod Mugutmal

D. Y. Patil College of Engineering, Pune

Abstract– The sentiment analysis is an emerging tool that analyzes data to generate insights on a specific topic. It serves government, corporations, and consumers by providing valuable information. In this project, machine learning-based approaches are applied to determine the sentiment of IMDb user reviews, evaluating whether the remarks are positive or negative. Multiple techniques such as Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and Support Vector Machines are employed to assess the polarity of sentiments in the dataset.

Keywords: Sentiment Analysis, Machine Learning (ML), Natural Language Processing (NLP), Movie Reviews

❖ INTRODUCTION

The sentiment analysis is the process of using natural language processing and computational linguistics to extract, identify and categorize different opinions expressed in text format. The sentiment analysis is the process o f using natural language processing and computational linguistics to extract, identify and categorize different opinions expressed in text format. The sentiment analysis is the process of using natural language processing and computational linguistics to extract, identify and categorize different opinions expressed in text format. Sentiment analysis has been investigated on levels ranging from document level, where the polarity of a sentiment of the whole document can be determined; to sentence level sentiment analysis, where polarity is determined sentence-wise; going deeper towards phrase-level sentiment analysis and finally aspect level sentiment analysis, where sentiments are selected on the basis of the aspects.

♦ MOTIVATION

The significance and importance of sentiment analysis are well-understood, as it aids businesses in comprehending their consumers' sentiments towards their brand or product. Automating this process allows stakeholders to make informed decisions based on the gathered data. Sentiment analysis can be defined as the utilization of methods and strategies to examine customer sentiments regarding a specific service or product. It involves automatically analyzing natural language occurrences, extracting important opinions or views, and categorizing them based on emotional attitude. The applications of sentiment analysis are diverse and can address various business needs, including customer satisfaction analysis. By conducting sentiment analysis, stakeholders can identify and promptly address any issues in real-time

❖ PROBLEM DEFINITION

The task of sentiment analysis is very crucial and needs detailed
attention. As discussed earlier, emoticons play a major role in deriving sentiments from one's review. Their presence or absence can have a major impact on model evaluation
parameters and hence are important for consideration. So the formulation of the problem statement is as follows - What is the effect of emoticons on sentiment analysis? This project utilizes various Machine Learning algorithms mentioned above to examine and compare the evaluation parameters and answer the problem statement.

❖ RELARED WORK

DATA COLLECTION

Data collection is an important step in sentiment analysis as it determines the approach to be used.

The collected data can be gathered from various sources such as web scraping, social media, news channels, E-commerce websites, forums, blogs, and weblogs.

2.Feature Extraction-

The problem accessed in this project is basically a binary classification i.e. all input reviews are needed to be correctly classified into two classes either positive or negative. Developing a model requires the identification of relevant features in the dataset such that a review can be partitioned into base words during the training and then appended into the feature vector. The basic technique involves the usage of unigrams, bigrams, or trigrams.

- **Methodology**

For the task of Sentiment Analysis, three main approaches are commonly used: Lexicon-based, Machine-based, and Hybrid approaches. The choice of approach depends on the type of data available. Machine learning techniques are suitable for structured data, while lexicon-based approaches are suitable for unstructured data, as they rely on predefined values associated with lexicons. When dealing with structured data, machine learning or deep learning approaches are employed, as they require minimal human effort and are suitable for automated sentiment analysis.

For semi-structured data, hybrid approaches are utilized. These approaches combine both lexicon-based and machine learning approaches to handle the different types of data effectively

- **Data Processing**

After selecting the data for the task of sentiment analysis, some preprocessing of data is required to make it model-ready with target of achieving maximum accuracy. Data after data selection is in raw form i.e. natural form. Data needs to go through various steps to be model-ready. Accuracy of a model depends directly on this step. All the feature extraction techniques mentioned above are employed on this step.

This project utilizes tf-idf vectorizer as a feature. TF-IDF vectorization involves calculating the TF-IDF score for every word in the corpus relative to that document (in this case, reviews) and then putting that information into a vector. This helps in transformation of the text into sensible representation of integers which may be used to fit machine learning algorithm for predictions. The term frequency is the number of occurrences of a specific term in a document. Term frequency indicates how important a specific term in a document. Document frequency is the number of documents containing a specific term signifying how common the term is.

Inverse document frequency (IDF) is the weight of a term, it aims to reduce the weight of a term if the term's occurrences are scattered throughout all the documents. It can be calculated using

$$Idfi=\log(n/dfi)$$

Finally ,tf-idf score can be calculated using

$$Wi,j = tfi,j * idfi$$

Where wi,j is tf-idf score for term i in review j, former term on R.H.S. is term frequency and latter term is inverse document frequency.

Standard preprocessing steps are employed further to convert raw data into meaningful and processible format. This is achieved in steps mentioned below. First step is done by removing all the HTML tags and markups from the reviews.

Second step has two versions, one involves removal of punctuation marks without saving emoticons for later appending and other involves saving emoticons for later

appending. This step is crucial for differentiating the affect of emoticons on sentimental analysis and differences on evaluation parameters is done on the basis of emoticon appending. Next step involves converting all letters to lowercase for text normalization, with and without saved emoticons appended. Then for each word in a review, tokenization and stemming is performed. Stemming is removal

of suffixes and prefixes from a word to bring it into its root form, while tokenization is breaking down of sentences into words. This is the final step of data preprocessing and now the data is ready to be employed into the models .

Finally we split our preprocessed data into training and testing sets keeping ratio of 7:3 with stop words removed paralelly. This is achieved by using a function train test split from the sklearn module for python. Training data is the data with which the models are trained. Testing data is the data on which testing is performed to evaluate all the parameters of evaluation of a model.

Logistic Regression

Logistic Regression is an example of supervised learning. It is used to forecast the probability of a binary event occuring. It is a type of statistical model often used for classification and predictive analysis. It estimates the probability of an event occuring based on a given dataset of independent variables. That is, it uses X_train for training and based on that it predicts the class label of X_test. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.In logistic regression, the probability of success divided by the probability of failure. It utilizes sigmoid function to classify the reviews into the positive and negative. Weight value is multiplied with the input value resulting into the values between 0 and 1. If the value is above 0.5, the review classifies as positive, otherwise as negative. Sklearn module provides the model and trains on X_train..

- **Decision Tree**

Decision Trees are a type of non-parametric supervised machine learning. They continuously split data based on specific parameters, which act as conditions or tests for creating subsets. This algorithm is suitable for binary classification problems, which is the requirement for this project. Decision trees consist of internal nodes and leaf nodes. They build classification or regression models using a tree structure. The dataset is recursively divided into smaller subsets, and a decision tree is incrementally developed. The resulting tree comprises decision nodes and leaf nodes. In this case, the leaf nodes represent positive and negative reviews, while the decision nodes are formed through machine learning techniques..

- **Random Forest**

Random Forest is a versatile and powerful supervised machine learning algorithm that utilizes ensemble learning. It overcomes the drawbacks of single decision trees by combining multiple decision trees into a forest. This algorithm improves performance by aggregating the outputs of multiple classifiers. Random Forest addresses overfitting by building decision trees on random feature subsets. The number of trees used directly affects the model's performance, with more trees generally leading to better results. The aggregated forecast is obtained by averaging the outputs in regression or selecting the majority vote in classification. Incorporating actual inputs enhances accuracy, and the algorithm works efficiently even on large datasets. The working process involves selecting random data points, building decision trees for those subsets, and repeating the process for a specified number of trees. For new data points, predictions are obtained from each decision tree, and the majority vote determines the category assignment. The only drawback of Random Forest is its suboptimal performance on regression tasks..

Naive Bayes-

Naive Bayes is a simple machine learning algorithm and a probabilistic model used for classification. It is based on Bayes' theorem, which states that the probability of event A can be determined by considering the conditional probability of its occurrence. of an event B given both events are –

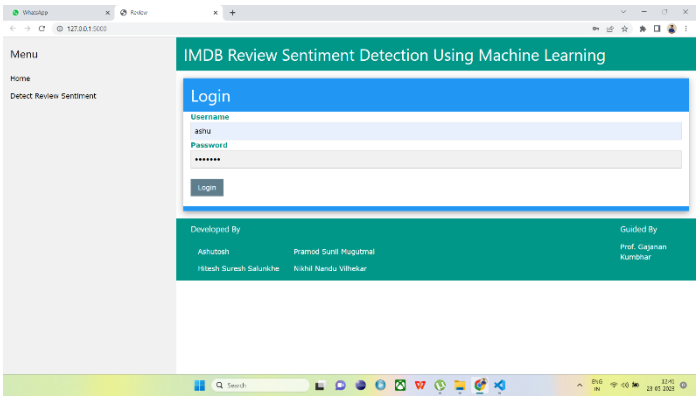
P(A/B) =P(B/A)p(A)/P(B)

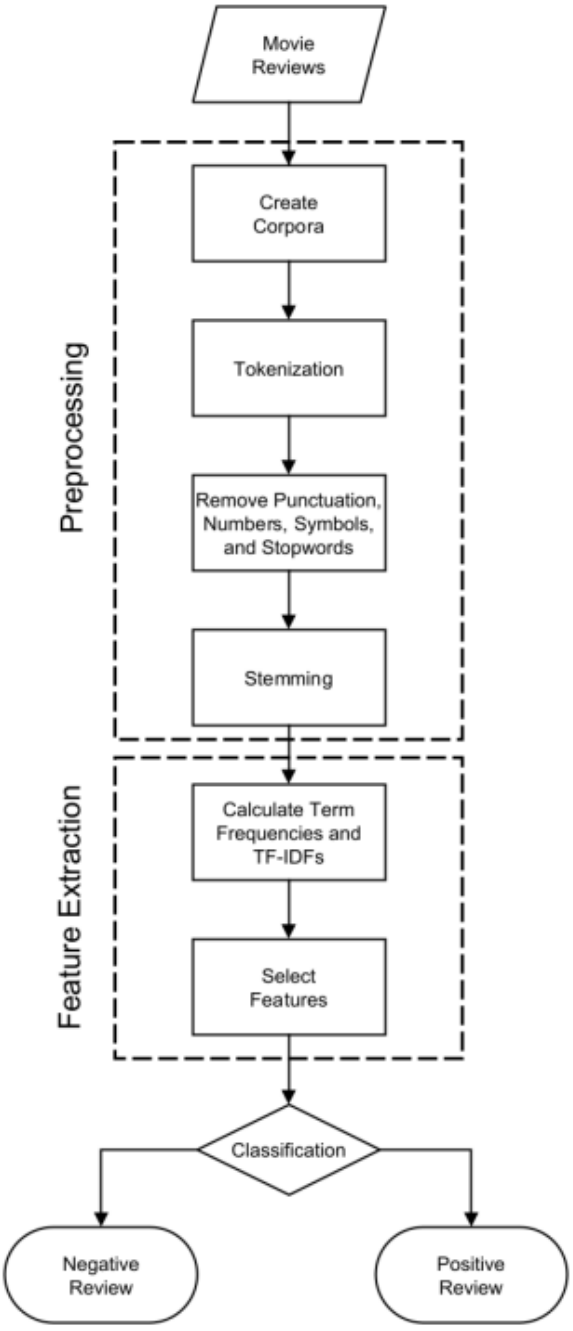
Where A and B are events and P(B) != 0

With regards to the given dataset, the above theorem can be applied as follows –

P(y/X) =P(X/y)P(y)/P (X)

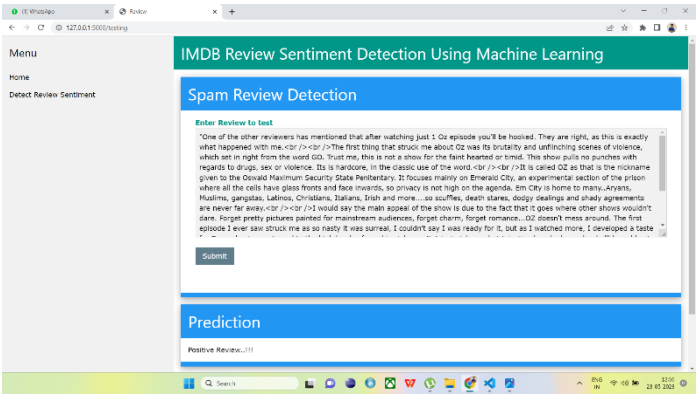
where, y is class variable and X is a dependent feature vector (of size n) where: X =(x1 ,x2,....) \
This algorithm is generally applied when training dataset is small.



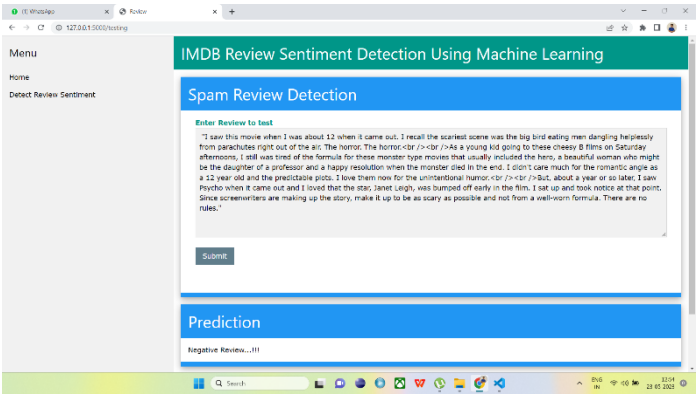


❖ Output

Home Page



Positive Review



Negative Review

❖ Support Vector Machine

Support Vector Machine (SVM) is a robust technique for classification and regression that maximizes predictive accuracy without overfitting. It is well-suited for analyzing data with a large number of predictor fields. SVM finds applications in various fields such as customer relationship management (CRM), image recognition, bioinformatics, text mining, intrusion detection, protein structure prediction, and speech recognition.

SVM works by mapping data to a high-dimensional feature space, enabling the categorization of data points even when they are not linearly separable. It finds a separator between categories and transforms the data to represent it as a hyperplane. Using these characteristics, SVM can predict the class label

(positive or negative) of new data records.

Evaluation Parameters

The majority of state-of-the-art sentiment analysis utilizes accuracy, F1 score, precision, and recall as performance metrics. In the review titled "Sentiment Analysis Using Deep Learning Architectures: A Review," recall and accuracy are highlighted as important metrics. The definitions of these metrics are as follows:

True Positive (TP): The number of positive reviews that have been correctly classified.
tive (TN): The number of negative reviews correctly classified as negative.

False Positive (FP): The number of incorrectly classified positive reviews.

False Negative (FN): The number of incorrectly classified negative reviews.

- **Precision**

Precision is is defined as the ratio of correctly classified positive samples to the total number of samples predicted as positive. This metric can be used to indicate the strength of the prediction. i.e., if a model has 100 percent precision, all the samples evaluated as positive are confidently positive. Precision = TP/(TP+FP).

- **Recall**

Recall Also known as sensitivity, recall is defined as
the ratio of actual positive instances to the total number of positive instances in the classification. It measures the misclassifications made by the model. Precision and recall have an inverse relationship, making it impossible to increase both at the same time. Recall is particularly useful when capturing a dominant class. The formula for recall is Recall = TP / (TP + FN).

- **F1 score**

The F1 score is the harmonic mean of Recall and Precision. It is a widely used metric, second only to Accuracy. The F1 score is employed when it is challenging to prioritize between Precision and Recall. It effectively balances the trade-off between recall and precision.

- **Accuracy**

Accuracy is the most commonly used metric in classification tasks. It defines the accuracy of the model by calculating the ratio of correct classifications to total predictions. Accuracy is a suitable metric for sentiment classification when dealing with a balanced dataset.

- **Confusion matrix**

A confusion matrix is a table commonly used to evaluate the effectiveness of a classification model on a known set of training test data values.

- **TF-IDF**

Term Frequency (TF) refers to the count of how many times a term appears in a document. Each document has a different length, so it is common for a term to appear more frequently in longer documents compared to shorter ones. To address this, the term frequency is often normalized by dividing it by the document length.

Conclusion

1. We achieved high accuracies using various machine learning methods, demonstrating their effectiveness in performing sentiment analysis. The Support Vector Machine Classifier exhibited the highest accuracy but
required longer training time. On the other hand, the Decision Tree algorithm had the lowest accuracy but required less training time. To enhance the results, additional steps in data preprocessing or alternative feature extraction methods could be explored beyond tf-idf vectorization. Notably, appending emoticons to the data resulted in increased model accuracy.

- **References**

Acheampong FA, Nunoo-Mensah H, Chen W (2021) Transformer models for text-based emotion detection: a review of BERT-based approaches. Adomavicius G, Kwon Y (2011) Improving aggregate recommendation diversity using ranking-based techniques. Ahmad S, Asghar MZ, Alotaibi FM, Awan I (2019) Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. Human Centric Comput Inf Sci 9(1):1–23. Bakar AA, Yaakub MR (2019) A feature selection techniques review in sentiment

analysis. Akhtar MS, Ekbal A, Cambria E (2020) How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble [application notes]. Akhtar N, Zubair N, Kumar A, Ahmad T (2017) Aspect-based sentiment-oriented summarization of hotel reviews. Al Amrani Y, Lazaar M, El Kadiri KE (2018) Random forest and support vector machine-based hybrid approach to sentiment analysis. Birjali M, Kasri M, Beni-Hssane A (2021) A comprehensive survey on sentiment analysis: approaches, challenges, and trends. Bhaskar J, Sruthi K, Nedungadi P (2015) Hybrid approach for emotion classification of audio conversation based on text and speech mining. Bhatia P, Ji Y, Eisenstein J (2015) Better document-level sentiment analysis from first discourse parsing. Calders T, Verwer S (2010) Three naive Bayes approaches for discrimination-free classification. Cambria E, Das D, Bandyopadhyay S, Feraco A (2017) Affective computing and sentiment analysis. Cao Q, Duan W, Gan Q (2011) Exploring determinants of vot