Sentimental Analysis of Movie Reviews Using NLP Techniques

Conference Paper · August 2023

DOI: 10.1109/ICIRCA57980.2023.10220783

6 authors, including:

Basi Reddy Avula
Mohan Babu University, Tirupati

Sivakumar Depuru
Sree Vidyanikethan Engineering College

Sakthivel Muthusamy
Sree Vidyanikethan Engineering College

Proceedings of the 5th International Conference on Inventive Research in Computing Applications (ICIRCA 2023)
DVD Part Number: CFP23N67-DVD; ISBN: 979-8-3503-2142-5

Sentimental Analysis of Movie Reviews using NLP Techniques

Basi Reddy A
Assistant professor, School of Computing, Mohan Babu University, Tirupati, A.P, India
basireddy.a@gmail.com

B.V. Suresh Reddy
Dept. of. CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P, India
sureshvenkat83@gmail.com

Anjana Devin Nandam
Dept. of. CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P, India
anjana.nandam@gmail.com

K Naresh
Assistant Professor, Sri Venkateswara College of engineering (autonomous), Tirupati, A.P, India
Nareshrpc123@gmail.com

Sivakumar Depuruu
School of Computing, Mohan Babu University, Tirupati, India
sivakumar.d@vidyanikethan.edu

M. Sakthivel
School of Computing, Mohan Babu University, Tirupati, A.P, India
sakthivel.m@vidyanikethan.edu

***Abstract*— The goal of natural language processing is to build machines that can comprehend text or speech input and respond to it in a way that is similar to how people do. To accomplish this, various Natural Language Processing (NLP) tasks are employed to analyze human language data, aiding computers in understanding the information it consumes. These tasks include sentimentanalysis, voice recognition, part-of-speech tagging, and natural language generation. Sentiment analysis that detects emotions such as happiness, frustration, rage, and sadness goes beyond polarity. To do this, advanced machine learning techniques or lexicon collections of words and the emotions arouse are used. In this study, the Internet Movie Database (IMDB) Reviews Dataset is subjected to sentiment analysis to determine if a user's review is favourable or unfavourable based on the supplied details. Utilizing data from the real world, the proposed approach's correctness is assessed.**

Keywords*— *Movie Reviews, Sentiment Analysis, NLP, Machine Learning.

I. INTRODUCTION

Movie reviews have become a popular way for critics to evaluate a film's overall quality and whether or not they recommend it. The purpose of these reviews is to help readers make informed decisions about whether to watch, rent, or buy a movie. A good review should provide enough detail about the film without giving away any key plot points or surprises. It gives the reader a sense of the movie's quality before watching it, as the authors delve into the nuances of the film [1, 19]. Additionally, there are brief but enjoyable reviews that give the reader an idea of whether a movie is worth their time. Although the ratings in the Internet Movie Database (IMDB) are determined by an objective algorithm, it is not regarded as beingabsolutely accurate. Instead, it gives viewers a quick method to see what other IMDB users around the world think of the movies that are featured on the website. IMDB users from around the world think of the movies that are featured on the website [2].

Before the advent of the internet, people often sought out opinions from acquaintances and friends when making decisions. However, with the rise of big data and internet usage, it is now possible to gather opinions from a large number of people on a particular subject. There are millions of individuals expressing their opinions on these platforms thanks to the rise in the number of websites, blogs, social media platforms, and reviews [3]. For various evaluation processes, it can be difficult and crucial to analyse these opinions and attitudes. Sentiment analysis is a method to extract meaningful information from a variety of source materials that include unrecognized viewpoints [4].

Sentiment analysis can be used to analyze a piece of text and determine the sentiment contained within it. The procedure entails segmenting the communication into topical parts and giving each topic a sentiment score. A movie's overall rating can be determined by this automatic approach, which also rates how favorable or unfavorablea movie review is. Basic pre-processing for text involves removing non-alphabetic characters, stop words, and changing all words to lowercase. Although text in IMDB reviews can be messy and include incorrect punctuation and spelling, it can be mapped to numerical representations for machine learning models [5]. The conclusions drawn from sentiment analysis can help in determining if the customer is angry, dislikes a specific aspect of a product, or is excited about a new release. By grouping the data into major categories, it is seen that the reactions are more positive, neutral, or negative.

II. PROBLEM STATEMENT

The primary intent of sentiment analysis is to assess all points of view and determine the general polarity of opinions on certain topics using various classification levels, such as positive or negative. Natural language processing (NLP) techniques are utilized in the sentimental analysis of the IMDb review datasets to determine whether the data is positive, negative, or neutral. Users can use this analysis to determine whether a movie is worth viewing or not. In today's world, people often rely on ratings and reviews given by previous users for movies or other products. Therefore, it is crucial to draw reliable conclusions from those reviews for the users' benefit. Users can make this choice with the help of a summary of all movie reviews without having to read every single one of them. Critics commonly evaluate and comment on movies on movie-rating websites, which aids viewers in determining whether the film is worth seeing.

III. OBJECTIVES

Sentiment analysis seeks to extract people's thoughts from unorganized reviews articles and classify them into sentiment classifications like positive, neutral, or negative, with the option of also taking into account "extremely positive" and "very negative". It also goes by the name "opinion mining" and extracts feelings and emotions from written materials using text mining and natural language processing (NLP). Social networking analysis also plays a big role in Sentiment analysis and Recommendation Systems, where users can express their thoughts on items through various online behaviours including commenting, liking, sharing, and publishing reviews or comments about products. These actions fall under the categories of user-user, user-community, and user-entity behaviour. User-community behaviour covers interaction in between user and a public, such as entering or quitting a community, to be a fan, or taking part in a community meeting. User-user behaviour includes communication between two users, like texting and playing games. User-entity behaviour is the creation of content, such as writing a blog post, an essay, a review, or uploading a photo to social media.

The objective is to create a model that can correctly categorise movie reviews as favourable or negative depending on their content. This can be helpful for a number of purposes, including gauging public opinion on a particular film or evaluating how reviews affect a film's box office performance.

Evaluation of the model's performance using multiple measures is another goal of the analysis. This will aid in identifying the model's advantages and disadvantages and offer suggestions on how to further enhance it. The overall overarching objective is to develop a model that can accurately classify movie reviews and provide data on the efficacy of the suggested system for sentiment analysis in the field of movie reviews.

IV. SCOPE

The scope of the study is to create a model that can correctly identify favourable and negative movie reviews. This method of analysis can be helpful in a variety of situations, such as when movie studios want to assess how audiences will respond to their films or when online merchants want to examine consumer feedback. The system can also be used to spot patterns in movie reviews, such as which elements of a movie (such as acting, plot, or special effects) are referenced in both positive and negative reviews most frequently.

The scope of the study includes:

1. Data collection: Gathering a large dataset of movie reviews from various sources such as websites, social media platforms, and online forums.
2. Data pre-processing: preparing the text data for the model's analysis by cleaning and converting it. This includes removing irrelevant information, punctuation, stop words, and other noise.
3. Feature extraction: To quantitatively represent the text data and give the model the ability to comprehend the relative value of various terms in the text, vectorizing the words using methods like TFIDF or bag-of-words is necessary.
4. Model selection: Choosing a suitable algorithm like support vector machine, multinomial naive bayes or Neural Networks for sentiment classification based on the size and complexity of the dataset.
5. Model training and evaluation: The dataset is divided into training and testing portions, and the model is predominantly trained on the training set. Metrics like precision, recall, accuracy, and F1 score are used to assess the model's performance onthe testing set.
6. Visualization and interpretation: Visualizing the classification results using tools such as confusion matrices, heatmaps, and WordCloud to gain insights into the sentiment of the movie reviews and identify trends and patterns.
7. Deployment: Integrating the sentiment analysis model into a web application, dashboard, or API for real-time monitoring and analysis of customer feedback.

The scope of the paper can be expanded to include other tasks such as topic modeling, sentiment intensity analysis, and aspect-based sentiment analysis based on the specific needs of the paper.

V. LITERATURE SURVEY

A. Movies Reviews Sentiment Analysis and Classification

The research is focused on building a sentiment analysis model for movie reviews. The model uses several techniques, such as tokenization, stemming, feature selection, and classification, to process and classify reviews as positive or negative. The researchers evaluated the model using eight different classifiers and five evaluation metrics on a real-world dataset [6].

According to the results, the Random Forest classifier performed the best among the eight classifiers, while Ripper Rule Learning performed the worst. This suggests that the Random Forest classifier is a good choice for sentiment analysis of movie reviews.

Overall, the research highlights the importance of sentiment analysis in enhancing the efficiency of products, such as movies. By classifying reviews as positive or negative, companies can gain insight into how to improve their products and better meet the needs of their customers [7].

B. Albert-based sentiment analysis of movie review

The article is discussing the importance of sentimentanalysis in analyzing movie reviews and how machine learning can help automate this process. The article mentions that sentiment analysis can extract topic information from text reviews and determine the overall polarity of the review.

The researchers in this article used the Albert model to build a sentiment analysis classifier and trained it using the "movie review dataset" from Stanford University. The Albert model achieved an accuracy of 89.05% when analyzing the sentiment of movie reviews, which is 3% higher than the traditional LSTM and GRU models [8].

This research suggests that machine learning models like the Albert model can be used to analyze movie reviews and extract sentiment information more efficiently and accurately than traditional methods. This can be helpful for movie producers and distributors to better understand the audience's opinions and improve the quality of their products accordingly [9].

C. An Improved Sentiment Analysis Of Online Movie Reviews Based On Clustering For Box-Office Prediction

The goal of this study appears to be to forecast a movie's box office success using sentiment analysis and machine learning techniques.

More online evaluations are being written as a result of the expansion of e-commerce, which may be a great source of knowledge for both buyers and sellers. The study shows that by including data from internet reviews, a simpler version of the sentiment- aware autoregressive model may be used to reliably forecast box office profits. To distinguish between positive and negative feelings, the technique employs fuzzy clustering and document-level sentiment analysis. The paper also discusses the usage of a Support Vector Machine (SVM) Classifier to forecast the trajectory of box office revenue based on review sentiment [10].

According to this strategy, sentiment analysis and machine learning techniques can be used to harness data from internet reviews for potential economic impact, such as forecasting a movie's box office performance. In conclusion, this study e mphasises the potential advantages of sentiment analysis and machine learning techniques for studying online reviews and forecasting the performance of goods and services. This might be helpful for companies and sectors that depend on consumer input to enhance their goods and services and d ecide wisely on marketing and sales tactics [11].

D. Sentiment Analysis of Movie Reviews: A Comparative Study between the Naive-Bayes Classifier and a Rule-based Approach

In order to categorise textual movie reviews as good or n egative, this paper relies on the use of sentiment analysis. In this study, the output of the Naive-Bayes algorithm and a rule-based approach that makes use of the AFINN-111 sentiment dictionary are compared.

To find and quantify subjective information in text data, sentiment analysis employs natural language processing, text analysis, and other computational tools. Popular machine learning algorithm the Naive-Bayes can be used for sentiment analysis. In this method, a model is trained on a labelled dataset of fav ourable and unfavourable reviews, and it is then used to forecast the sentiment of future reviews [12].

On the other hand, the rule-based approach uses a sentiment dictionary or a pre-defined set of rules to ascertain the sentiment of a review. As it includes a list of words with corresponding sentiment scores that may be used to categorise reviews, the AFINN-111 sentiment dictionary is a popular option for sentiment analysis. In order to evaluate whether strategy is more successful for sentiment analysis of movie reviews, the findings of the Rule-Based strategy and the Naive-Bayes algorithm are compared in this study using the AFINN-111 sentiment dictionary. The study's conclusions can be applied to organisations and sectors that depend on customer feedback, including movie studios and internet streaming services, to enhance the goods and services they provide [14] .

VI. EXISTING SYSTEM

Existing systems that use NLP to analyse sentiment in IMDB review data are numerous. Feature extraction, data preparation, and model training are c common procedures in the sentiment study of IMDB movie re views using NLP approaches. Cleaning up the raw text data, getting rid of any extraneous information like stop words and punctuation, and turning the text into a numerical representation like a bag-of-words or word embeddings are all parts of data preparation [15]. NLP methods like sentiment lexicons, sentiment polarity scores, and sentiment classification models are used to provide features for feature extraction that accurately reflect the sentiment expressed in the text. Finally, a machine learning model or neural network is trained on the extracted information to forecast the sentiment of a particular review. The accuracy of the model is evaluated using metrics like precision, recall, and F1 score, and it is then improved [16].

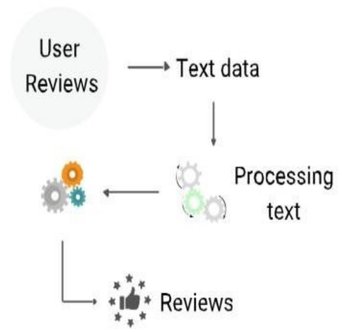

Fig 1: Natural Language Processing

1. Data Preprocessing

To preprocess text for analysis, the following steps are commonly taken:

- Eliminate irrelevant information such as punctuation, numbers, and stop words.
- Tokenize the text into individual words.
- Convert the words into their base form using stemming orlemmatization.
- Generate a mathematical representation of the tokenized text, such as a bag-of-words or word embeddings.

2. Feature extraction Sentiment lexicons:

The overall sentiment polarity of the text is calculated using a pre-defined list of words with corresponding sentiment polarity r ratings. Sentiment polarity scores: The sentiment polarity of each word is predicted using an existing model, which is then applied to the entire text to determine its overall sentiment polarity.

Sentiment classification models: A machine learning model is trained on a tagged dataset of movie reviews to forecast the sentiment of a particular review.

3. Model Training

After preprocessing, split the data into training and testing sets. A machine learning model is trained using the training set. The model is trained using the features that were retrieved. Analyze the model's performance on the testing set using metric measures once it has been trained.

4. Model Fine-Tuning

Adjust the hyper parameters of the model to enhance its performance. Iterate the process of training and evaluation until desirable results are achieved.

Some of these systems are:

i. *Rule-based systems:*

A collection of pre-established rules are utilised in a rule-based sentiment analysis system to ascertain the sentiment of a given text. These rules are often developed by subject matter specialists and are founded on their knowledge of the language and the subject matter of the analysis [17].

The rules can be simple or complex, and can take into account various factors such as the presence of certain words, phrases, or grammatical structures that are indicative of a particular sentiment. For example, a rule might state that the presence of words such as "love," "great," and "amazing" indicate a positive sentiment, while the presence of words such as "hate," "terrible," and "disappointing" indicate a negative sentiment.

One advantage of a rule-based system is that it can be more interpretable than other approaches such as machine learning. It allows for explicit control over which features are used to determine sentiment, and can be easier to modify and fine-tune as needed.

However, a potential limitation of rule-based systems is that they may not be as accurate or flexible as machine learning-based approaches, particularly when dealing with complex or ambiguous language. Additionally, creating and maintaining a set of rules can be time-consuming and require significant expertise in the domain being analyzed.

ii. Statistical models:

Statistical methods in sentiment analysis involve using mathematical models and techniques to analyze and interpret the sentiment of text data. These techniques frequently make use of machine learning al gorithms, which are trained on textual datasets with labels to find links and patterns between words and attitudes.

One common statistical method used in sentiment analysis is the bag-of-words approach, which involves representing each piece of text as a collection of individual words, ignoring their order and context. Each word is given aweight based on how frequently it appears in the text and how significant it is to the sentiment being studied. These weights are then used to calculate an overall sentiment score for the text.

Sentiment analysis lexicons, which are collections of words that have been carefully annotated with sentiment polarity scores, are another statistical technique used in sentiment analysis. The sentiment scores of each word can be added together to determine the overall sentiment polarity of a passage of text [18].

Overall, because of their precision and capacity for handling vast volumes of data, statistical approaches are frequently used in sentiment analysis. However, they might need a lot of computer power and might not always be able to deal with how complicated s poken language and context are.

iii. Deep Learning Models:

Due to their capacity to recognise intricate patterns in data, deep learning models are becoming more and more common in sentiment analysis. They are a particular class of machine learning model made up of numerous artificial neuronal layers that have been taught to recognise hierarchical representations of the input data. The use of deep learning models in the area of sentiment analysis have demonstrated encouraging results in correctly predicting the sentiment of a given text. They are able to discover intricate patterns and connections between words and sentences in a text because they can be trained on vast volumes of labelled data.

The recurrent neural network (RNN), which excels at analysing sequential data, such as text, is a popular model for deep learning applied to sentiment analysis. RNNs process each word individually while keeping an internal memory of the words that came before it, allowing them to capture the context and relationships between words in a sentence.

Convolutional neural networks are yet another well-liked deep learning model utilised in sentiment analysis. (CNN). CNNs have been modified for text classification applications like sentiment analysis since they are excellent at picture recognition. CNNs are particularly good at picking up local aspects of the input data, like n-grams or particular sentences that express a certain emotion.

iv. *Hybrid Systems:*

To increase the precision and robustness of sentiment analysis, hybrid systems integrate the advantages of many methodologies. These systems can leverage the advantages of rule-based systems, statistical methods, and deep learning models.

In a hybrid system, multiple models are combined to generate a final sentiment prediction. For example, a rule-based system may be used to classify simple and explicit sentiments, while a deep learning model may be used to identify more complex and nuanced sentiments. Statistical methods may also be used to refine the sentiment prediction and improve its accuracy.

Using an ensemble learning methodology, such as bagging or boosting, is a well-liked method for developing a hybrid system. In bagging, various models are individually trained on various subsets of data, and their predictions are then merged to get a final prediction. In boosting, the mathematical models are trained progressively, with each model paying particular attention to the data points that the prior model incorrectly identified.

Another approach for building a hybrid system is to use a stacked architecture, where the output of one model is used as input to another model. To recognise explicit sentiment, for instance, a rule-based system may be employed. The output of this system may then be utilised as input to a deep learning model that identifies more complicated and nuanced sentiment.

Overall, hybrid systems have been shown to outperform individual systems in sentiment analysis, especially in cases where the sentiment is complex and nuanced. By combining the strengths of different approaches, hybrid systems can provide more accurate and robust sentiment predictions.

Various systems have been developed for sentiment analysis on reviews data with varying levels of accuracy. Theselection of a system depends on factors such as the size and characteristics of the dataset, the desired level of accuracy, and available computational resources.

VII. Proposed System

The Multinomial Naive Bayes algorithm-based proposed system for sentiment analysis is a text classification model that foretells whether a given text will be positive or negative. The technology is made to assist companies and organisations in analysing consumer comments and reviews, posts on social media, and other text data to comprehend customer attitude towards their goods and services. The system represents the incoming text data as a vector of word counts using the bag-of-words method before classifying it into positive or negative sentiment using the Multinomial Naive Bayes methodology. The bag-of-words approach treats each word as a unique feature, disregarding the context and word order of the text. With this method, the data representation is made simpler, and the algorithm may concentrate on the key phrases that are most likely to convey the text's mood.

The system weighs the significance of each word in the text. Based on a word's frequency in the text and rarity across all corpus documents, the TF-IDF algorithm assigns each word a score. This method gives terms that appear more frequently in the text but less frequently across all documents a higher weight, indicating that these words are more distinctive and useful for the classification of sentiment.

1. Data Preprocessing: Data preparation is the initial step in any machine learning that aims to clean up and transform raw data into a format that machine learning algorithms can use. The movie reviews are subjected to a number of data cleaning procedures in order to get them ready for sentiment analysis.

a) HTML Stripping: The movie reviews are initially processed to remove any HTML tags using the BeautifulSoup library in Python. This step is essential to eliminate any HTML code that may be present in the reviews and could introduce noise or irrelevant information.

b) Square Bracket Removal: The square brackets in the movie reviews are removed using regular expressions to eliminate any text enclosed in square brackets, which may contain noise or irrelevant information that does not contribute to sentiment analysis.

c) Special Character Removal: A function is defined to remove special characters from the movie reviews using regular expressions. This helps remove any non-alphanumeric characters from the data that could not be useful for sentiment analysis or could cause noise.

d) Text Normalization: The movie reviews are further normalized by applying text normalization techniques such as stemming and stopword removal. Stemming is performed using the PorterStemmer algorithm from the NLTK library, which reduces words to their root or base form. Stopwords, which are common words such as "the" and "is", are removed from the reviews using a predefined list of English stopwords from the NLTK library. This step helps in reducing the dimensionality of the data and focuses on the meaningful words that contribute to sentiment analysis.

2. Feature Extraction: The technique of feature extraction involves turning unstructured data, such text, into numerical features that can be fed into machine learning algorithms. Two well-known feature extraction methods are applied in this suggested system: Bag of Words (BoW) and Term Frequency-Inverse Document Frequency. (TF-IDF).

a) Bag of Words (BoW): The CountVectorizer class from the sklearn library is used to implement BoW in the given code. BoW represents text as a collection of words and their frequencies in a document. It creates a numerical matrix where each row, which represents a document, and each column, which represents a word, have values that represent the frequency of each word. The machine learning models use this matrix as their input data.

b) Term Frequency-Inverse Document Frequency (TF-IDF): The TfidfVectorizer class from the sklearn library is used to implement TF-IDF. The TF-IDF algorithm uses a numerical matrix to represent text. Each row corresponds to a document, and each column to a word. The values in the matrix represent the importance of each word in a given document based on the frequency of the word in the document and the inverse frequency of the word over the whole dataset. In order to apply weights to words, TF-IDF considers both the term frequency (TF) and inverse document frequency (IDF), providing greater weight to words that are common in a text but uncommon over the entire dataset. The machine learning models use this matrix as their input data.

3. Model Training and Evaluation: After data preprocessing and feature extraction, the next step in the proposed system is to train and evaluate machine learning models for sentiment analysis. Logistic Regression, Naive Bayes, and Support Vector Machines are three widely used machine learning methods.

a) Logistic Regression: Implementing logistic regression involves using the LogisticRegression class from the Sklearn package. A binary classification approach called logistic regression models the association between input features and a binary output variable. (positive or negative sentiment in this case). The BoW or TF-IDF matrices are used as input features and their corresponding labels are used as output targets to train the model on the training data. After the model has been trained, its performance is assessed using measures like accuracy, precision, recall, and F1-score on the test data.

b) Naive Bayes: The MultinomialNB class from the sklearn library is used to implement Naive Bayes. Naive Bayes is a probabilistic algorithm that assumes the independence of features, and it is commonly used for text classification tasks. The model is trained and evaluated in a similar manner as logistic regression, using the BoW or TF-IDF matrices as input features and their corresponding labels as output targets.

c) Support Vector Machines (SVM): SVM is implemented by the SVC class from the sklearn library. SVM is a potent binary classification technique that locates the optimum hyperplane for classifying the data. The BoW or TF-IDF matrices are used as input features and their corresponding labels are output targets in a manner similar to logistic regression and Naive Bayes for training and evaluation of the model.

4. Model Selection and Hyperparameter Tuning: Model selection and hyperparameter tweaking are essential processes in the proposed approach to maximise the effectiveness of the machine learning models.. The proposed

system selects a model and tunes hyperparameters using the GridSearchCV class from the Sklearn package. GridSearchCV conducts a thorough search over a predetermined parameter grid and chooses the optimal set of hyperparameters that maximises the model's performance.This helps in finding the optimal hyperparameters for each machine learning algorithm, such as the regularization strength for logistic regression or the kernel type for SVM, to achieve the best possible performance on the movie review dataset.

5. Model Deployment: It is feasible to deploy the top-performing model for real-world applications after choosing it and fine-tuning its hyperparameters. The model can be saved as a serialized file using the joblib library in Python, which allows for easy deployment and integration into production systems. The deployed model can then be used to predict the sentiment of movie reviews in real-time or batch processing, depending on the application requirements.

The suggested system for sentiment analysis offers a thorough method for categorising movie reviews as positive or negative based on the sentiment indicated in the text. The system involves several critical steps, including data preprocessing, feature extraction, model training and evaluation, model selection and hyperparameter tuning, and model deployment. The proposed system aims to achieve accurate and reliable sentiment analysis on movie reviews by utilising well-known machine learning algorithms like logistic regression, Naive Bayes, and SVM, as well as techniques like Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction. The system can be further optimized by fine-tuning hyperparameters and incorporating additional techniques, such as word embeddings or deep learning algorithms, based on the specific requirements of the application. Overall, the proposed system holds great potential for various applications in the field of sentiment analysis and can contribute to improved decision-making, customer satisfaction, and business insights.

The proposed system has the potential to be further optimized by fine-tuning hyperparameters, incorporating additional techniques such as word embeddings or deep learning algorithms, and adapting to specific application requirements. It can be used in various applications such as sentiment analysis for customer reviews, social media sentiment analysis, brand monitoring, and market research, among others.

In conclusion, the proposed system for sentiment analysis using machine learning techniques is a comprehensive approach that involves multiple steps and techniques to accurately classify movie reviews as positive or negative based on their sentiment. It has the potential to provide valuable insights and decision-making support in various domains.

VIII. Algorithm Used

A. *Load and Preprocess Data:*

- Load the movie review dataset or any other dataset containing text data and their corresponding sentiment labels (positive or negative).
- Carry out data preprocessing operations such punctuation removal, text conversion to lowercase, stop word elimination, and stemming or lemmatization to break down words into their root forms.

B. *Extract Features:*

- Create numerical features from the preprocessed text data that may be fed into the machine learning algorithms.
- Word embeddings like Word2Vec or GloVe as well as TF-IDF (Term Frequency-Inverse Document Frequency) representation are common strategies for feature extraction.

C. *Split Data:*

- Split the preprocessed and feature-extracted data into training and testing sets.
- Typically, a random split of the data is employed, such as 80/20 or 70/30, with a larger amount being used for training and a smaller portion being used for testing.

D. *Implement Logistic Regression:*

- Create a logistic regression model using the retrieved features and training data.
- Use a suitable optimization algorithm such as gradient descent or any other optimization technique to learn the coefficients (parameters) of the logistic regression model.
- Tune hyperparameters such as regularization strength (if applicable) and learning rate to optimize the model's performance.
- Assess the logistic regression model's performance using the testing data's accuracy, precision, recall, and F1-score metrics.

E. *Implement Naive Bayes:*

- Train a Naive Bayes model using the training data and the extracted features.
- Implement the Naive Bayes algorithm, which involves calculating the probabilities of each feature occurring in each class and combining them using Bayes' theorem to calculate the conditional probabilities of the class given the input features.
- Assess the Naive Bayes model's performance using the testing data's accuracy, precision, recall, and F1-score metrics.

F. *Implement SVM:*

- Train an SVM model using the training data and the extracted features.
- If necessary, translate the data into a higher-dimensional space using a suitable kernel function (such as a linear, polynomial, or radial basis function).

- Identify the optimum hyperplane (decision boundary) for separating the data points into various classifications.
- To enhance the performance of the model, adjust hyperparameters such regularisation strength, kernel parameters, and cost parameter.
- Utilize performance indicators from the testing data, such as accuracy, precision, recall, and F1-score, to assess the SVM model.

G. Model Selection and Hyperparameter Tuning:

- Use techniques such as GridSearchCV to perform model selection and hyperparameter tuning.
- GridSearchCV involves trying different combinations of hyperparameter values for each algorithm and selecting the one that gives the best performance.
- Use cross-validation to get a more robust estimate of model performance and avoid overfitting.

H. Choose Best-performing Model:

- Compare the performance of the logistic regression, Naive Bayes, and SVM models using the evaluated metrics.
- Choose the best-performing model based on the performance metrics and the specific requirements of the sentiment analysis task.

I. Deploy and Test Model:

- Once the best-performing model is chosen, deploy it in a production environment or any application where sentiment analysis is required.
- Test the model with new, unseen text data to evaluate its performance in a real-world scenario.
- Monitor and update the model periodically to ensure its accuracy and effectiveness over time.

IX. RESULTS

The accuracy score and classification report for TF-IDF features will be the algorithm's final output. These findings will give us a better notion of how accurate the algorithm is at predicting the tone of movie reviews.

The percentage of reviews in the testing set that were correctly categorised is used to calculate the classifier's accuracy. The system computes precision, recall, and F1-score in addition to accuracy, which are metrics frequently used to assess the effectiveness of a binary classification model.

Out of all the reviews that the classifier projected would be positive, precision is the percentage of reviews that were correctly classified as positive. Recall quantifies the percentage of all positively classified reviews in the testing set that were correctly identified. The harmonic mean of accuracy and recall is the F1-score, which provides a single metric to evaluate the trade-off between precision and memory.

These metrics are thoroughly broken down for both positive and negative assessments in the classification report that the classification system generates. The report includes the precision, recall, F1-score, and support for both classes. The number of reviews in each class in the testing set is represented by the support.

This system generate classification reports for two different models - one that uses the bag of words approach and another that uses the TF-IDF approach. A classification report provides information about the accuracy and precision of a classification model. It includes various metrics such as precision, recall, F1-score, and support.

The **classification_report()** function takes three arguments - the true sentiment values, the predictedsentiment values, and the target names for the sentiment categories. In this case, the target names are "Positive" and "Negative".

The first report is for the bag of words model, while the second report is for the TF-IDF model. These reports provide an evaluation of how well the models have performed in classifying the sentiment of the movie reviews.

	precision	recall	f1-score	support
Positive	0.75	0.76	0.75	4993
Negative	0.75	0.75	0.75	5007
accuracy			0.75	10000
macro avg	0.75	0.75	0.75	10000
weighted avg	0.75	0.75	0.75	10000

	precision	recall	f1-score	support
Positive	0.75	0.76	0.75	4993
Negative	0.75	0.74	0.75	5007
accuracy			0.75	10000
macro avg	0.75	0.75	0.75	10000
weighted avg	0.75	0.75	0.75	10000

Table 1: Classification Report

The **WordCloud** function from the **wordcloud** library to create a word cloud visualization of the most frequently occurring words in the respective reviews. The **WordCloud** function takes in various parameters such as the width and height of the image, maximum number of words to include in the cloud, and minimum font size.

The resulting word clouds are displayed using **imshow** function from **matplotlib** library. The **interpolation** parameter is used to specify the interpolation method used to display the image.

Overall, the purpose of the code is to visually display the most commonly occurring words in the positive and negative movie reviews, and to gain insights into the sentiment expressed in the reviews.

The important terms that are frequently and extensively used in favourable reviews may be seen in the word cloud made from the words found in the positive reviews.

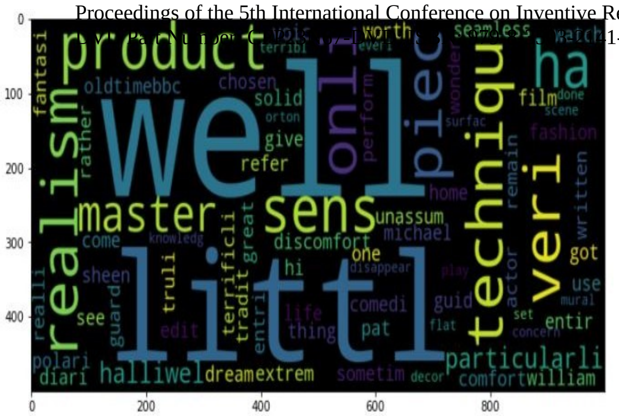

Fig 2: Positive Reviews wordcloud

The word cloud created from the negative review words can also provide some light on the key terms that are frequently and extensively utilised in negative reviews.

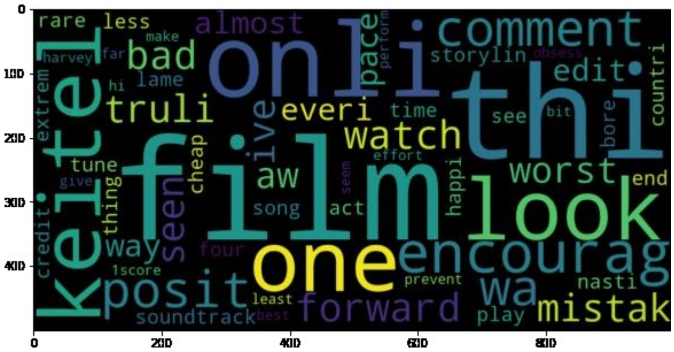

Fig 2: Negative Reviews wordcloud

Overall, the IMDB dataset shows that the Multinomial Naive Bayes classifier performs effectively, reaching better accuracy. Positive and negative reviews have good precision, recall, and F1-scores, demonstrating that the classifier can successfully separate the two classes. These findings imply that the model is applicable to sentiment analysis tasks on comparable datasets. It is critical to keep in mind that the performance of the model can vary depending on the precise dataset and the task at hand.

CONCLUSION

The dataset received preprocessing utilising techniques like stemming, stop word removal, special character removal, square bracket removal, and HTML stripping. The feature extraction algorithms used vectorization, also known as term frequency-inverse document frequency (TF-IDF). The TF-IDF feature extraction approach produces an accuracy score and a classification report using the Multinomial Naive Bayes classification algorithm.

In conclusion, the project's goal was to categorise movie reviews as either good or unfavourable. To do this, the proposed system used a variety of data pretreatment procedures, feature extraction approaches, and classification algorithms.

REFERENCES

[1] Nagamma P, Pruthvi H.R, Nisha K.K and Shwetha N H. An Improved Sentiment Analysis Of Online Movie Reviews Based On Clustering For Box-Office Prediction. International Conference on Computing, Communication and Automation (ICCCA2015).

[2] Mais Yasen, Sara Tedmori. Movies Reviews Sentiment Analysis and Classification.2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology(JEEIT).

[3] Zhongxiang Ding,Yali Qi,Deping Lin. Albert-based sentiment analysis of movie review.2021 4th International Conference on Advanced Electronic Mterials, Computers and Software Engineering(AEMCSE).

[4] Vihaan Nama,Dr.Vinay Hegde,Dr.B.Satish Babu.Sentiment Analysis of Movie Reviews: A Comparative Study between the Naïve-Bayes Classifier and a Rule-based Approach

[5] Depuru, Sivakumar, Anjana Nandam, P. A. Ramesh, M. Saktivel, and K. Amala. "Human Emotion Recognition System Using Deep Learning Technique." Journal of Pharmaceutical Negative Results 13, no. 4 (2022): 1031-1035.

[6] Sivanantham, S., V. Krishnamoorthy, D. Karthikeyan, M. Sakthivel, V. Mohanraj, and V. Akshaya. "Web Hazard Identification and Detection Using Deep Learning-A Comparative Study." Journal of Pharmaceutical Negative Results 13, no. 4 (2022): 1052-1060.

[7] Depuru, Sivakumar, Anjana Nandam, S. Sivanantham, K. Amala, V. Akshaya, and M. Saktivel. "Convolutional Neural Network based Human Emotion Recognition System: A Deep Learning Approach." In 2022 Smart Technologies, Communication and Robotics (STCR), pp. 1-4. IEEE, 2022.

[8] Saravanakumar, V., K. G. Suma, M. Sakthivel, K. S. Kannan, and M. Kavitha. "Segmentation of hyperspectral satellite image based on classical clustering method." Int J Pure Appl Math 118, no. 9 (2018): 813-820.

[9] Sakthivel, M., S. Sivanantham, R. Kamalraj, and V. Krishnamoorthy. "An Analysis of Machine Learning Depend on Q-MIND for Defencing the Distributed Denial of Service Attack on Software Defined Network." International Journal of Early Childhood Special Education 14, no. 05 (2022): 3769-3776.

[10] Bandana, Rachana. "Sentiment analysis of movie reviews using heterogeneous features." In 2018 2nd International conference on electronics, materials engineering & nano-technology (IEMENTech), pp. 1-4. IEEE, 2018.

[11] Rehman, Anwar Ur, Ahmad Kamran Malik, Basit Raza, and Waqar Ali. "A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis." Multimedia Tools and Applications 78 (2019): 26597-26613.

[12] Qaisar, Saeed Mian. "Sentiment analysis of IMDb movie reviews using long short-term memory." In 2020 2nd International Conference on Computer and Information Sciences (ICCIS), pp. 1-4. IEEE, 2020.

[13] S. Depuru, P. Hari, P. Suhaas, S. R. Basha, R. Girish and P. K. Raju, "A Machine Learning based Malware Classification Framework," 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2023, pp. 1138-1143, doi: 10.1109/ICSSIT55814.2023.10060914.

[14] Tan, Kian Long, Chin Poo Lee, and Kian Ming Lim. "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research." Applied Sciences, 13, no. 7, 2023.

[15] William, P., Anurag Shrivastava, Premanand S. Chauhan, Mudasir Raja, Sudhir Baijnath Ojha, and Keshav Kumar. "Natural Language processing implementation for sentiment analysis on tweets." In Mobile Radio Communications and 5G Networks: Proceedings of Third MRCN 2022, pp. 317-327. Singapore: Springer Nature Singapore, 2023.

[16] Bordoloi, Monali, and Saroj Kumar Biswas. "Sentiment analysis: A survey on design framework, applications and future scopes." Artificial Intelligence Review, 2023.

[17] Namitha, Seelam, Pallerla Sanjan, Nerella Chaithra Reddy, Yechuri Srikar, Hariharan Shanmugasundaram, and Bhanu Prasad Andraju. "Sentiment Analysis: Current State and Future Research Perspectives." In 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1115-1119. IEEE, 2023.

[18] Yadav, Vandana, Parul Verma, and Vinodini Katiyar. "Long short term memory (LSTM) model for sentiment analysis in social data for e-commerce products reviews in Hindi languages." International Journal of Information Technology 15, no. 2, 2023.

[19] A. B. Reddy and R. Y. R. Kumar, "Performance and Security Analysis in Cloud Using Drops and T-Coloring Methods," 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India, 2022, pp. 1-7, doi: 10.1109/ICERECT56837.2022.10060014.