IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

TWITTER SENTIMENT ANALYSIS USING TEXT CLASSIFICATION

[1]M.SAI PRASAD, [2]R JEGADEESAN, [3]B.ANNAPURNA, [4]E.RAVALIKA,[5]M.BHAVANA

[1,2,3]Final Year DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
[1]Asst. Professor,[2] Professor, Jyothishmathi Institute of Technology and Science
Karimnagar, Telangana.

ABSTRACT

Twitter, with its vast user base and diverse opinions, serves as a dynamic platform influencing trends and business strategies. Analyzing sentiments expressed in tweets is crucial for enhancing customer service, driving website traffic, and optimizing marketing campaigns. However, manual analysis of this massive textual data is arduous and time-consuming.

Sentiment analysis, an automated process leveraging AI, identifies positive and negative opinions within text. Our objective is to analyze sentiment in tweets using the Sentiment140 dataset. To achieve this, we're developing a machine learning pipeline employing three classifiers: Logistic Regression, Bernoulli Naive Bayes, and Support Vector Machines (SVM). Additionally, we'll utilize Term Frequency-Inverse Document Frequency (TF-IDF) and Long Short-Term Memory (LSTM) techniques.

The performance of these classifiers will be evaluated based on accuracy and F1 Scores. Subsequently, we'll compare the models to determine their effectiveness in sentiment analysis.

Key Words:CNN,RNN, Multinomial Naive Bayes(MNB),SVM,Logistic Regression(LR),Sentiment Analysis

I INTRODUCTION

The acquisition of Twitter by Elon Musk underscores the platform's significance in the 21st century. With the rapid growth of the internet, social media platforms like Instagram, Meta, Twitter, WhatsApp,andReddit have surged in popularity. Twitter, specifically, functions as a microblogging platform where individuals share their thoughts, ideas, insights, and emotions on various topics, including products and current events. In today's landscape, it's essential for companies to gauge user sentiment following product launches, and for governments to understand public reception to their initiatives. Twitter sentiment analysis presents numerous opportunities in this regard. By analyzing tweets in real-time and deciphering the sentiment behind each message, businesses and governments gain deeper insights into public opinion.

One of the primary applications of Twitter sentiment analysis in business is social media monitoring. Maintaining a positive online reputation is crucial for brands, as negative reviews or mentions can harm their image and bottom line. Twitter sentiment analysis enables organizations to monitor social media conversations about their products

or services, identifying and addressing any negative sentiment promptly. Additionally, it provides valuable insights into what customers appreciate and dislike about a brand, aiding decision-making processes. For instance, a tweet expressing satisfaction with Amazon's fast shipping highlights the importance of this aspect to customers.Our project focuses on sentiment analysis of tweets sourced from Twitter using a Long Short-Term Memory (LSTM) classification model. Additionally, we aim to assess the sentiment of tweets from the Sentiment140 dataset by constructing a machine learning pipeline. This pipeline incorporates three classifiers (Logistic Regression, Bernoulli Naive Bayes, and SVM) along with Term Frequency-Inverse Document Frequency (TFIDF) for feature extraction. We evaluate the performance of these classifiers using accuracy and F1 Scores as metrics. This allows us to compare their effectiveness in accurately classifying tweet sentiments.

II LITERATURE SURVEY

In the paper [1],The review paper provides a comprehensive overview of the foundational knowledge required to conduct sentiment analysis on Twitter. Sentiment analysis, considered a subset of text mining and natural language processing, encompasses various dimensions that are explored in this paper. These include different levels of sentiment analysis, diverse approaches to conducting it, methodologies employed, and the extraction of features from text, all of which contribute to understanding sentiment in textual data.Twitter, being a microblogging platform, presents unique challenges for sentiment analysis. The paper delves into the process of extracting tweets, preprocessing them, and ultimately analyzing their sentiment. It categorizes the paper as a review paper and highlights keywords such as Classifier, Lexicons, Machine Learning, Opinion Mining, Sentiment Analysis, Twitter, and Text Mining.Furthermore, the paper classifies models used for sentiment analysis into probabilistic classifiers, rule-based classifiers, linear classifiers, and decision trees, providing a comprehensive framework for conducting sentiment analysis on Twitter data.The main aim in research paper [2]The paper utilizes two existing datasets for sentiment analysis: the "Sentiment140" dataset from Stanford University, containing 1.6 million tweets, and another dataset sourced from "Crowdflower's Data for Everyone library" with 13,870 entries. Both datasets are pre-categorized based on the sentiments expressed in them.Several sentiment classifiers, including Textblob, Sentiwordnet, Multinomial Naive Bayes (MNB), Logistic Regression (LR), Support Vector Machines (SVM), and Recurrent Neural Network (RNN) Classifier, are applied to these datasets. The paper compares the results obtained from these classifiers, which classify tweets as either positive or negative based on their sentiment.In addition to individual machine learning approaches, the paper explores an ensemble approach involving MNB, LR, and SVM classifiers on the datasets. The performance of this ensemble method is then compared with the results obtained from individual classifiers.Furthermore, the trained models can be utilized for sentiment prediction on new data, enabling the application of sentiment analysis to real-time or future datasets.In the paper[3] Our proposal suggests integrating various feature extraction methods, including emoticons, exclamation and question marks, word gazetteers, and unigrams, to enhance the accuracy of sentiment classification systems. This paper conducts an empirical comparison of six supervised classification algorithms to evaluate their performance in sentiment classification.Inthis paper[4] focuses on analyzing Twitter posts related to electronic products such as mobile phones and laptops using a Machine Learning approach. By conducting

sentiment analysis within a specific domain, we aim to understand the impact of domain-specific information on sentiment classification. We introduce a novel feature vector for categorizing tweets as positive or negative and extracting people's opinions about products.The accuracies, strengths, weakness and applications are also stated which helps is proper choice of the algorithms for further.

III PROPOSED METHOD

In our approach, we use a hybrid feature extraction method that combines Convolutional Neural Network(CNN) with Principal Component Analysis(PCA).This technique aims to transform raw data into numerical features, effectively preserving crucial information inherent in the original dataset. CNN, renowned for its capability to recognize patterns within data, is employed to extract intricate features from the input. Meanwhile, PCA, a dimensionality reduction technique, further refines these features by identifying and retaining the most significant components, thus reducing computational complexity and enhancing model efficiency. By integrating CNN with PCA, our system achieves a synergistic effect, leveraging the strengths of both approaches to extract comprehensive and representative features from the input data. This hybrid feature extraction process ensures that the transformed features accurately capture the underlying patterns and nuances present in the original dataset, facilitating robust and effective analysis for tasks such as sentiment analysis on platforms like Twitter.

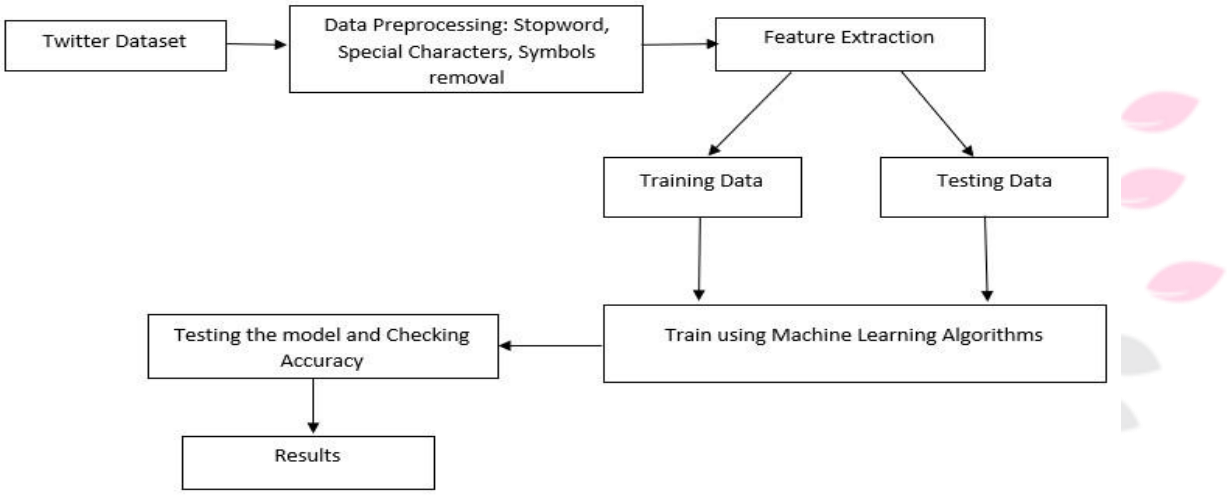


Fig 1: System Architecture

Bernoulli Naive Bayes

Bernoulli Naive Bayes is a widely utilized technique in Twitter sentiment analysis for text classification. Initially, a dataset comprising tweets and their corresponding sentiment labels (positive, negative, or neutral) is gathered. These tweets undergo preprocessing steps like tokenization, removing stop words, and stemming to prepare them for analysis. Then, each tweet is represented as a binary feature vector, where the presence or absence of words or features is indicated. These binary feature vectors, along with their sentiment labels, serve as input for training the Bernoulli Naive Bayes classifier. During training, the algorithm learns the likelihood of observing certain features (words) given each sentiment label. Once trained, the model can predict the sentiment of new tweets by calculating the probability of each sentiment label given the observed features in the tweet, using Bayes' theorem. This allows for real-time sentiment analysis on Twitter, providing insights into public opinion, brand perception, and other relevant topics. Bernoulli Naive Bayes offers a simple yet effective approach to sentiment analysis, particularly suitable for handling textual data in social media platforms like Twitter due to its computational efficiency and ability to handle high-dimensional feature spaces.

Support Vector Machines

Support Vector Machines (SVMs) play a crucial role in Twitter sentiment analysis through text classification. Initially, a dataset containing tweets and their associated sentiment labels (such as positive, negative, or neutral) is compiled. Following this, the tweets undergo preprocessing steps like tokenization, stop word removal, and

stemming to prepare them for analysis. Each tweet is then converted into a numerical feature vector, representing various text features. These feature vectors, along with their sentiment labels, serve as input for training the SVM model. During the training process, the SVM algorithm learns to classify tweets into different sentiment categories by finding the optimal hyperplane that maximizes the margin between different classes while minimizing classification errors. Once trained, the SVM model can predict the sentiment of new tweets by placing them in the appropriate sentiment category based on their feature vectors. This predictive capability enables real-time sentiment analysis on Twitter, providing insights into public opinion, brand perception, and other relevant topics. SVMs offer a robust and effective approach to sentiment analysis in the dynamic and noisy environment of social media platforms like Twitter, making them a valuable tool for understanding and interpreting large-scale textual data.

logistic regression

In Twitter sentiment analysis using text classification, logistic regression serves as a key algorithmic tool. Initially, a dataset comprising tweets and their corresponding sentiment labels is collected. These sentiments typically include positive, negative, or neutral classifications. The tweets then undergo preprocessing steps such as tokenization, removing stop words, and stemming to prepare them for analysis. Subsequently, each tweet is transformed into a feature vector, where numerical values represent various text features. These feature vectors, along with their sentiment labels, are used to train the logistic regression model. During training, the model learns the relationship between the features extracted from the text and the sentiment labels assigned to the tweets. Once trained, the model can predict the sentiment of new tweets based on their feature vectors. This predictive capability allows for the real-time analysis of sentiment on Twitter, offering valuable insights into public opinion, brand perception, and other relevant topics. Overall, logistic regression provides a robust and interpretable framework for sentiment analysis in the dynamic environment of social media platforms like Twitter.

IV RESULT

After assessing all the models, the following Results can be drawn

• Accuracy: Logistic Regression outperforms SVM, which in turn outperforms Bernoulli Naive Bayes in terms of delicacy.

• F1- score:

• For class 0: Bernoulli Naive Bayes< SVM< Logistic Regression

• For class 1: Bernoulli Naive Bayes< SVM< Logistic Regression

• AUC Score: All three models have the same ROC- AUC score. Grounded on these findings, it can be concluded that Logistic Regression is the most suitable model for the dataset anatomized.

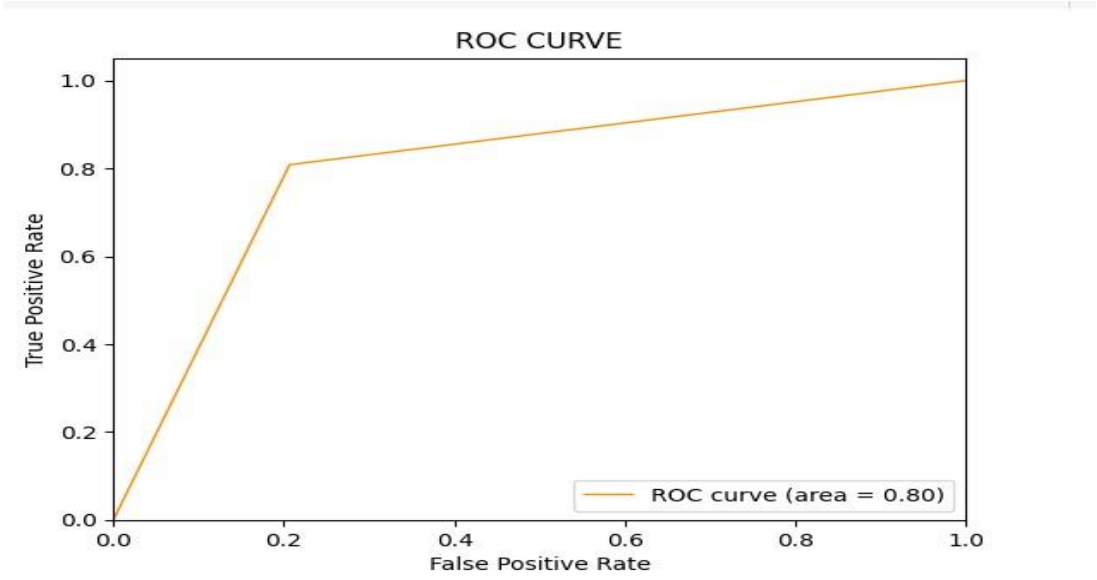
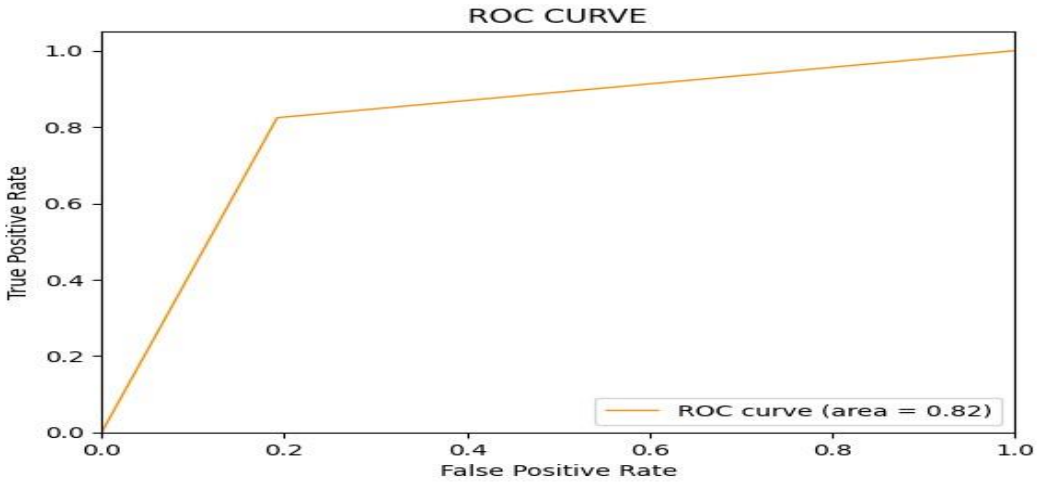


Fig 2:Model-1 Accuracy



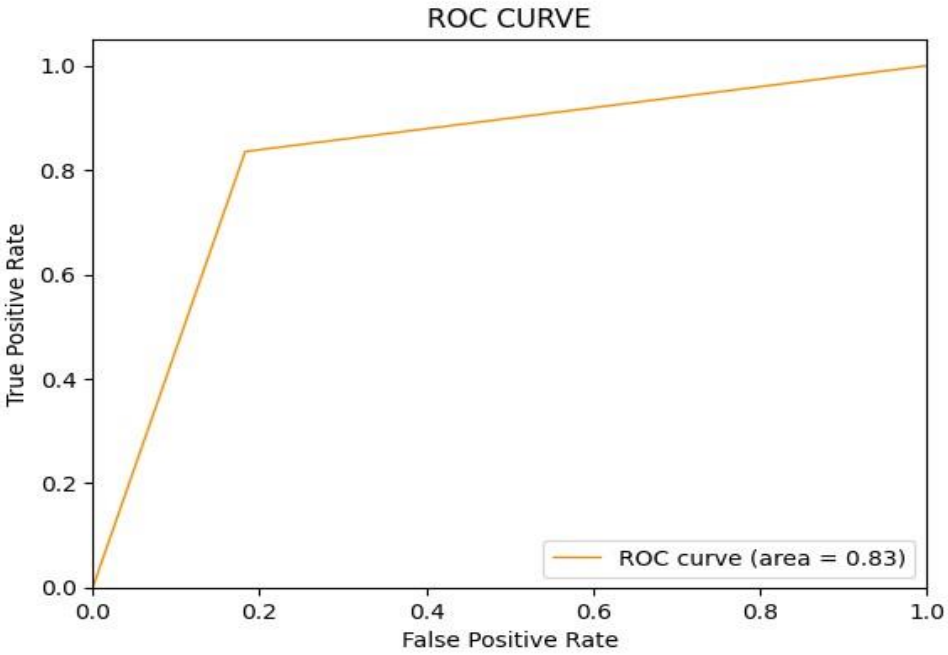


Fig 3:Model-2 Accuracy

Fig 4:Model-3 Accuracy

V CONCLUSION

In summary, we trained three supervised classification algorithms on the dataset: Logistic Regression, Support Vector Machines (SVM), Bernoulli Naive Bayes, Classifier. These classifiers were chosen based on their popularity and effectiveness in sentiment analysis tasks. We utilized the Sentiment140 dataset from Stanford University, consisting of 1.6 million tweets. Dataset were pre-categorized based on the sentiments expressed in the tweets.The performance of each classifier was evaluated using metrics such as accuracy, F1-score, and area under the ROC curve (AUC). Accuracy measures the proportion of correctly classified instances, while the F1-score balances precision and recall for each class. The AUC score reflects the model's ability to distinguish between positive and negative sentiments.

Upon evaluating the models, several key observations were made:

- Logistic Regression consistently outperformed SVM and Bernoulli Naive Bayes in terms of accuracy.
- The F1-scores for both positive and negative classes were highest for Logistic Regression, followed by SVM and Bernoulli Naive Bayes.
- All three models achieved similar AUC scores, indicating comparable performance in distinguishing between positive and negative sentiments.

Based on these findings, we concluded that Logistic Regression is the most suitable model for sentiment analysis of the given dataset. Its superior performance across multiple metrics underscores its effectiveness in accurately classifying tweets as positive or negative sentiments.In conclusion, our study highlights the importance of selecting appropriate features and classifiers for sentiment analysis tasks. By combining various feature extraction techniques and evaluating multiple classifiers, we can gain valuable insights into public opinion on social media platforms like Twitter. Additionally, our findings emphasize the significance of Logistic Regression as a reliable model for sentiment analysis, with potential applications in marketing, customer feedback analysis, and brand reputation management.

REFERENCES

[1]Pawar, Kishori & Shrishrimal, P & Deshmukh, Ratnadeep. (2015).Twitter Sentiment Analysis: A Review. Ijser. 6.

[2]Dr. Priyanka Harjule, Astha Gurjar, Harshita Seth, Priya Thakur.(2020).Text Classification on Twitter Data.

[4] R Jegadeesan,A. Beno,S. P. Manikandan,D. S. Naga Malleswara Rao,Bharath Kumar Narukullapati,5T. RajeshKumar,Batyrkhan Omarov,Areda Batu, "Stable Route Selection for Adaptive Packet Transmission in 5G-BasedMobile Communications", "Wireless Communications and Mobile Computing 2022 "Research Article | OpenAccess Volume 2022 | Article ID 8009105 | https://doi.org/10.1155/2022/8009105.

[5] M. Akshitha, R Jegadeesan, G. Akshaya, P. Akhilac, M.Pavan Kalyan, G.Sindhusha, 2021 & June, "Covid-19Future Forecasting Using Supervised Machine Learning Models", Zeichen Journal, Volume 7, Issue 6, Page No.257-269, ISSN No: 0932-4747. DOI:15.10089.ZJ.2021.V7I6.285311.2425 (UGC Care Group II Journal)

[6] PerukaPriyavarshini, R Jegadeesan, Thatla Vaishnavi, KampellySahithi, Boga Shivani, P.Balakishan, 2021 &June, "Cyber Money Laundering Detection Using Machine Learning", Zeichen Journal, Volume 7, Issue 6, 2021,Page No.231-238,ISSN No: 0932-4747. DOI:15.10089.ZJ.2021.V7I6.285311.2422 (UGC Care Group II Journal)

[7] R Jegadeesan, Dava Srinivas, N Umapathi, G Karthick, N Venkateswaran "Personal Healthcare Chatbot ForMedical Suggestions Using Artificial Intelligence And Machine Learning", European Chemical Bulletin, Eur.Chem. Bull. 2023, 12 (S3), 6004 – 6012, DOI: 10.31838/ecb/2023.12.s3.670. (Scopus)

[8]]Ajay Deshwal & Sudhir Kumar Sharma.(2021).Twitter Sentiment Analysis using Various Classification Algorithms.

[9]Neethu M S& Rajasree R.(2019).Sentiment Analysis in Twitter using Machine Learning Techniques.