SENTIMENT ANALYSIS OF TWITTER DATA

MINI PROJECT REPORT
Machine Learning [MTE 4073]

Submitted by

Isha Harish (Reg. No. 200929098)
Diya Parekh (Reg. No. 200929176)

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

DEPARTMENT OF MECHATRONICS

NOVEMBER 2023

ABSTRACT

This mini-project explores the dynamic landscape of sentiment analysis on Twitter data, recognizing its significance in understanding public opinions and emotions in the contemporary digital era. The primary objective is to employ advanced AI techniques for sentiment analysis, utilizing tools such as wordclouds for intuitive visualization. The study employs classic machine learning models—Naive Bayes, Logistic Regression, and Gradient Boosting Classifier—to analyze sentiment patterns. The methodology involves preprocessing Twitter data, feature extraction, model training, and evaluation. Results indicate the effectiveness of these classifiers in discerning sentiment nuances within the tweets. The significance lies in the practical applications of these models, offering valuable insights into public sentiments on diverse topics, making this mini-project a pertinent contribution to the evolving field of sentiment analysis on social media platforms.

Contents

Chapter 1: Introduction

1.1

Sentiment analysis, a burgeoning field within natural language processing, delves into the automated interpretation and understanding of sentiments expressed in textual data. In the era of social media dominance, platforms like Twitter serve as a rich source of real-time, diverse, and voluminous text data reflecting public opinions and emotions. This mini-project aims to navigate this digital landscape, leveraging advanced artificial intelligence (AI) techniques for sentiment analysis on Twitter data. The objective is to unravel the intricate fabric of sentiments embedded in tweets, paving the way for a nuanced understanding of the collective mood in the virtual public sphere.

Present-Day Scenario:

In the contemporary landscape, the importance of sentiment analysis has never been more pronounced. Social media platforms, particularly Twitter, have become dynamic arenas where individuals express sentiments ranging from joy to outrage, shaping and reflecting societal narratives. The ability to harness this torrent of textual data for sentiment analysis is pivotal for businesses, policymakers, and researchers seeking to gauge public sentiments, track trends, and respond effectively to the evolving discourse.

Motivation for the Project:

The motivation behind this mini-project stems from the realization that in the age of information overload, distilling meaningful insights from vast volumes of text data is paramount. Understanding the sentiments permeating through social media conversations is crucial for various applications, from business intelligence and brand perception analysis to political sentiment tracking. The project's motivation lies in the potential to contribute valuable tools and methodologies to the sentiment analysis toolkit, addressing the growing need for nuanced sentiment interpretation in the context of Twitter data.

Objectives:

- Implement advanced AI techniques, including Naive Bayes, Logistic Regression, and Gradient Boosting Classifier, for sentiment analysis on Twitter data.
- Develop a robust preprocessing pipeline to handle the unique characteristics of Twitter text, including hashtags, mentions, and emoticons.
- Create visually insightful representations of sentiment patterns using wordclouds, enhancing interpretability and user engagement with the analysis results.

- Evaluate the performance of each sentiment analysis model using appropriate metrics such as accuracy, precision, recall, and F1 score.

- Compare and contrast the performance of Naive Bayes, Logistic Regression, and Gradient Boosting Classifier in the context of sentiment analysis on Twitter data.

- Provide a comprehensive discussion of the results, highlighting the strengths and limitations of each model and the practical implications for real-world applications.

As we embark on this endeavor, each objective is designed to contribute to the overarching goal of advancing our understanding of sentiments in the digital realm and providing actionable insights for applications across diverse domains.

1.2

Sl. No.	Student Name	Registration Number	Individual objective
1	Isha Harish	200929098	Literature review, Editing tabular data,Clean up for stopwords,punctuations and stemming
2	Diya Parekh	200929176	Literature review ,using different AI models,Evaluating them using Confusion

Chapter 2: Literature Review and Theoretical Background

1.1 Present State / Recent Developments:

Recent developments in sentiment analysis have been characterized by a surge in interest in harnessing machine learning models to discern nuanced sentiments in Twitter data. Advanced techniques, including deep learning models, have gained traction, yet classic algorithms like Naive Bayes, Gradient Boosting Classifier, and Logistic Regression continue to be relevant due to their interpretability and efficiency. Ensemble learning, represented by Gradient Boosting Classifier, has particularly seen widespread adoption owing to its ability to capture complex relationships within data.

1.2 Naive Bayes Classifier:

Naive Bayes relies on Bayes' theorem and assumes independence between features, making it computationally efficient and interpretable. The theoretical foundation lies in probability theory, where the likelihood of a class given certain features is calculated based on the product of individual feature probabilities. Despite its simplicity, Naive Bayes has shown effectiveness in text classification tasks, including sentiment analysis.The mathematical derivation involves Bayes' theorem, where the posterior probability is proportional to the product of the prior probability and the likelihood. The assumption of feature independence simplifies calculations, leading to a straightforward estimation process.

Multinomial and Bernoulli Variants:

Researchers have explored different variants of Naive Bayes, such as Multinomial and Bernoulli Naive Bayes, each suitable for specific types of data. Multinomial Naive Bayes, for instance, is commonly employed in text classification tasks, while Bernoulli Naive Bayes is suitable for binary feature data.

1.3 Logistic Regression:

Logistic Regression is a classic statistical model that predicts the probability of a binary outcome. The model is based on the logistic function, mapping predictions to a probability range. The theoretical discussion involves the likelihood function and maximum likelihood estimation. Logistic Regression is valued for its simplicity, interpretability, and robustness in capturing linear relationships.

Logistic Regression involves the logistic function, linking the odds of the dependent variable to a linear combination of the independent variables. The model parameters are estimated through maximum likelihood estimation, and the resulting logistic curve represents the probability of the positive class.

Foundations of Logistic Regression in Sentiment Analysis:

Logistic regression is well-suited for binary classification tasks, making it particularly relevant for sentiment analysis, where the goal is often to determine whether a piece of text expresses positive or negative sentiment. The logistic function employed in this regression model maps predictions to a probability range (0 to 1), facilitating the interpretation of sentiment probabilities.

Feature Representation and Model Training:

Studies have explored various feature representations for sentiment analysis using logistic regression, including bag-of-words, n-grams, and word embeddings. The model is trained on labeled datasets, learning the relationship between features and sentiment labels. Regularization techniques, such as L1 and L2 regularization, have been investigated to prevent overfitting and enhance generalization.

Real-world Applications and Comparative Studies:

Logistic regression's applicability extends to real-world scenarios, such as product reviews, social media sentiment, and customer feedback analysis. Comparative studies with other machine learning algorithms, including support vector machines and deep learning models, have been conducted to benchmark logistic regression's performance in sentiment analysis tasks.

1.4 Gradient Boosting Classifier:

Gradient Boosting is an ensemble learning method that builds a series of weak learners sequentially, with each tree correcting the errors of its predecessor. The theoretical underpinning involves minimizing a loss function, and the final model is a weighted sum of the weak learners. Gradient Boosting excels in capturing non-linear relationships in data, making it suitable for sentiment analysis on Twitter where nuanced sentiments prevail. Mathematical derivations of Gradient Boosting involve the computation of gradients and the iterative minimization of a loss function. The ensemble of weak learners is constructed by optimizing the weighted sum of residuals, resulting in a powerful predictive model.

Handling Non-linearity in Sentiment Analysis:

Sentiment analysis often involves handling non-linear relationships between text features and sentiment labels. Gradient Boosting excels in capturing such non-linearity, allowing the model to discern subtle contextual cues and dependencies that may be challenging for linear models.

Feature Importance and Interpretability:

Gradient Boosting provides a measure of feature importance, aiding in the interpretation of sentiment analysis models. This is crucial for understanding the linguistic aspects that contribute to sentiment predictions, facilitating insights into the key factors influencing positive or negative sentiment expressions.

Addressing Imbalanced Datasets:

Sentiment analysis datasets frequently exhibit class imbalance, with one sentiment class dominating over others. Gradient Boosting classifiers can be adapted to handle imbalanced datasets by assigning different weights to each class or by utilizing techniques like SMOTE (Synthetic Minority Over-sampling Technique).

Comparison with Other Models:

Comparative studies between Gradient Boosting classifiers and other machine learning models, such as Random Forests or Support Vector Machines, have been conducted in sentiment analysis tasks. These studies aim to benchmark the performance of Gradient Boosting and understand its strengths and limitations in comparison to alternative approaches.

Chapter 3: Problem Definition and Objectives

Problem Highlights:

The burgeoning use of social media, particularly Twitter, has led to an exponential increase in unstructured textual data, presenting significant challenges for effective sentiment analysis. The dynamic nature of Twitter, coupled with linguistic variations, noise, and the need for nuanced sentiment interpretation, underscores the necessity for advanced methodologies. The present mini-project endeavors to address these challenges by refining sentiment analysis techniques tailored to the unique intricacies of Twitter data.

Objectives:

1. **Design and Implement Preprocessing Techniques:**
   design and implement preprocessing techniques tailored for Twitter data.
   Employ stemming, remove stopwords and punctuation to enhance the efficiency of sentiment analysis.

2. **Utilize Wordcloud for Visual Representation:**
   incorporate wordclouds as a visual representation tool.
   Visualize sentiment patterns in an intuitive and insightful manner to aid interpretation.

3. **Implement Count Vectorization:**
   implement count vectorization as a crucial step in feature extraction.
   Transform the textual data into a numerical format suitable for machine learning algorithms.

4. **Apply Naive Bayes Classifier:**
   apply the Naive Bayes classifier for sentiment analysis.
   Utilize the probabilistic model to make predictions based on the extracted features.

5. **Apply Logistic Regression:**
   apply Logistic Regression for sentiment analysis.
   Leverage the logistic function to model the probability of sentiment classes.

6. **Apply Gradient Boosting Classifier:**
   apply the Gradient Boosting Classifier for sentiment analysis.
   Utilize ensemble learning to capture complex relationships within the Twitter data.

7. **Evaluate Model Performance using Confusion Matrix:**
   evaluate model performance using confusion matrix metrics.
   Assess the true positives, true negatives, false positives, and false negatives for each classifier.

8. **Compute Precision, Recall, and F1 Score:**
   compute precision, recall, and F1 score for each sentiment analysis model.
   Utilize these metrics to evaluate the models' effectiveness in capturing sentiment nuances.

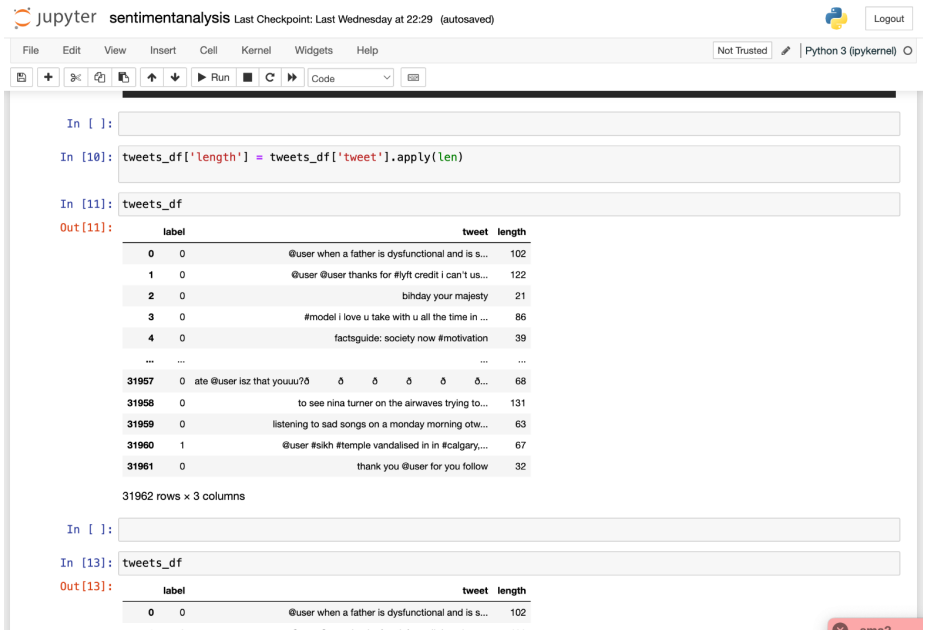
9. **Comparative Analysis of Models:**
   conduct a comparative analysis of Naive Bayes, Logistic Regression, and Gradient Boosting Classifier.
   Identify the strengths and weaknesses of each model to inform model selection.

10. **Provide Comprehensive Results Discussion:**
    provide a comprehensive discussion of the results.
    Emphasize the practical implications of findings, offering insights for real-world applications.

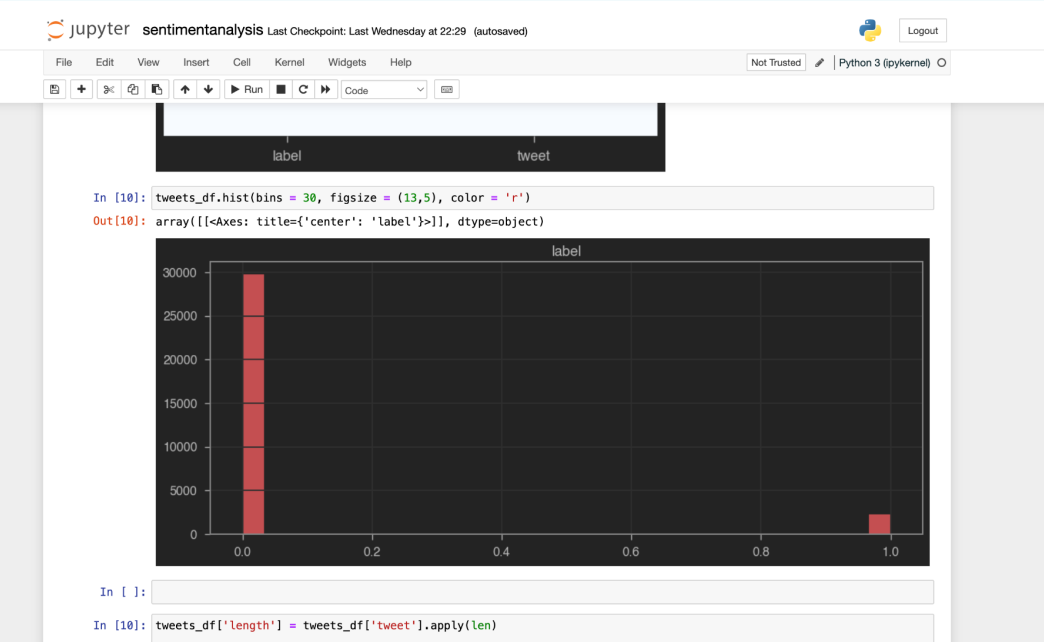
By delineating these objectives, the mini-project aims to systematically address the complexities inherent in sentiment analysis on Twitter data, utilizing a combination of preprocessing techniques, advanced algorithms, and rigorous evaluation measures to enhance the accuracy and interpretability of sentiment predictions.

Chapter 4: Methodology

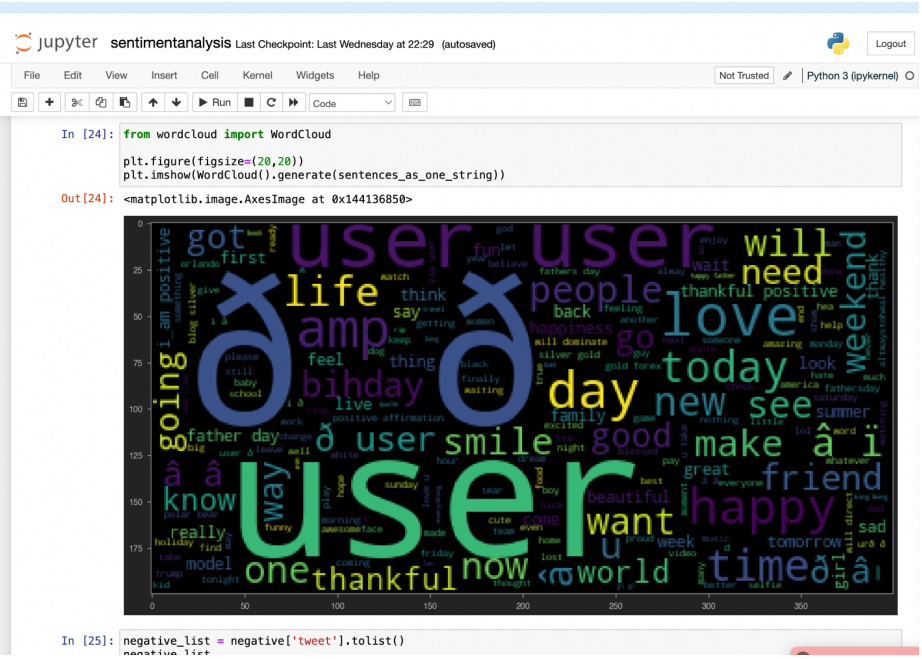
Initially, we examine the tabular dataset and make adjustments to the columns based on specific requirements. In our dataset, we excluded the "id" column and introduced a new column labeled "length," furnishing us with the character count of each tweet.We got the dataset from kaggle.



Employing graphical representations, such as histograms, allows us to visualize the distribution of values and identify any potential null values that may necessitate removal.



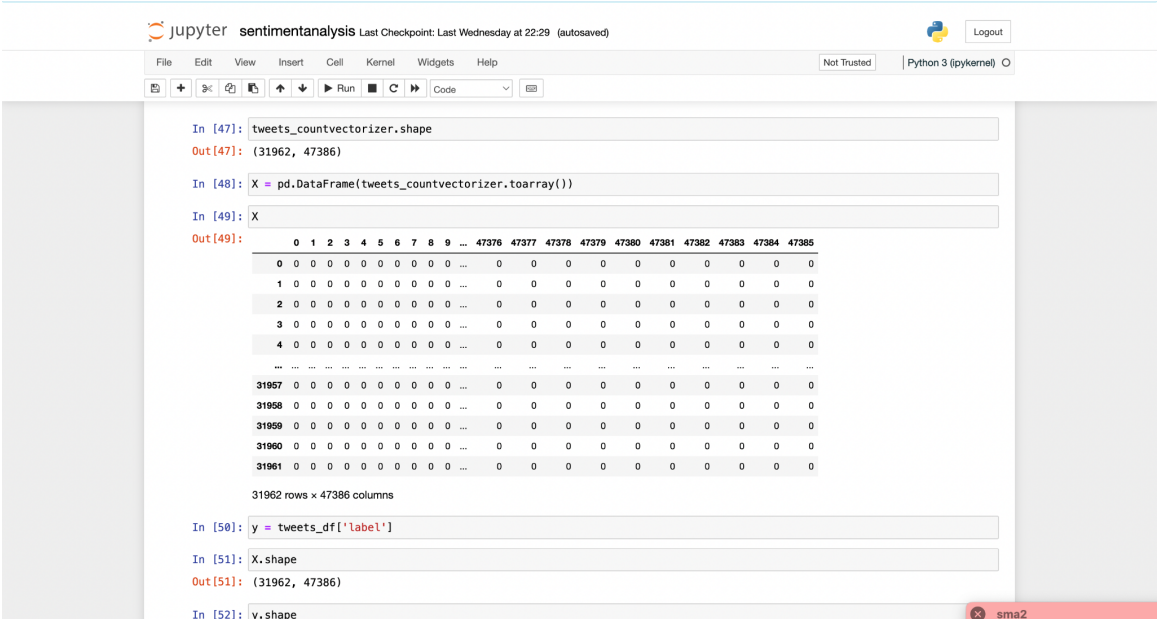
Identify positive and negative tweets within the training data, amalgamate them into a unified string, and utilize this composite string to generate a word cloud for visualization purposes.



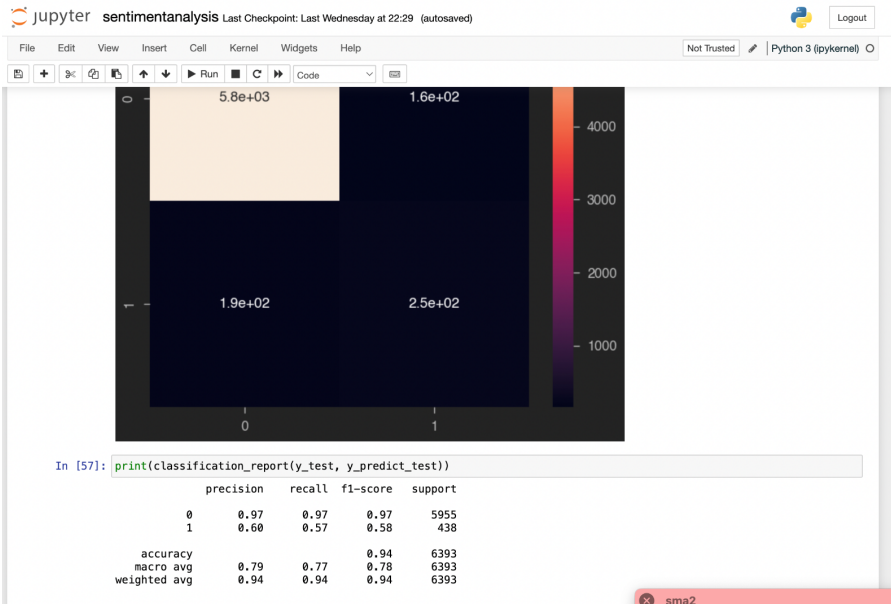
Eliminate all punctuation, conduct stemming, and exclude stopwords to enhance the cleanliness of the dataset, thereby optimizing it for superior results.

Following these preprocessing steps, proceed with count vectorization. This technique in natural language processing transforms a set of text documents into a matrix reflecting the

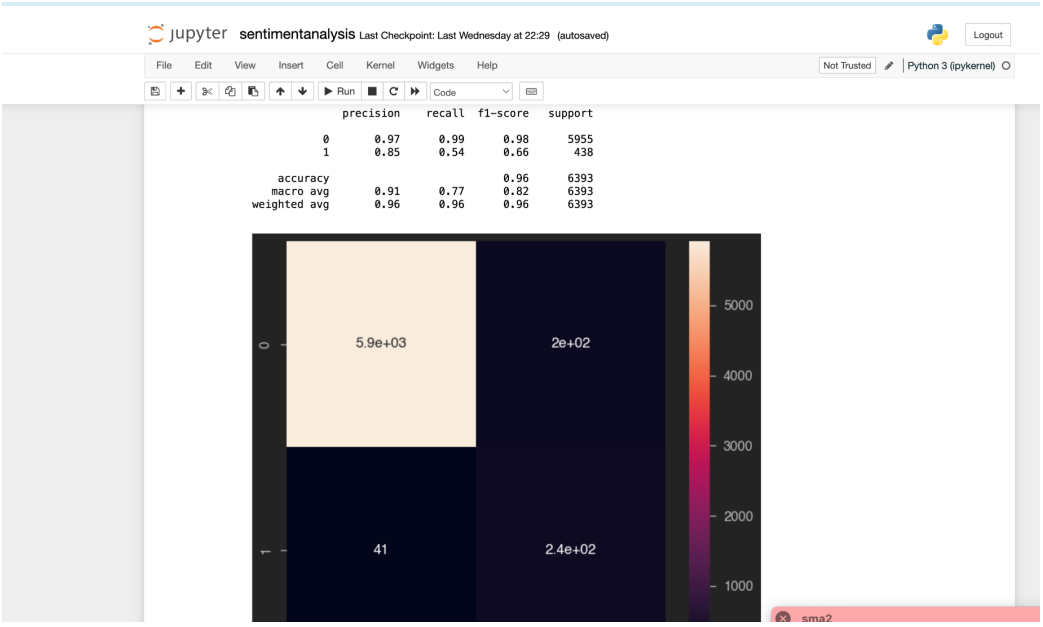
frequency of terms. Each document is represented as a vector, wherein each element signifies the occurrence count of a particular word within the document. Count vectorization is a pivotal step in extracting meaningful features from the text data, providing a structured representation conducive to subsequent machine learning analyses.



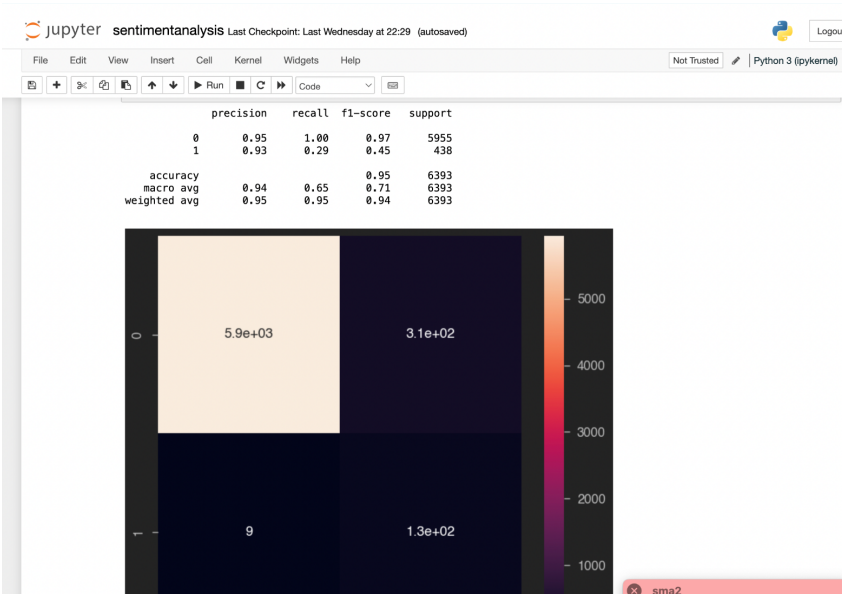
After this we use naive Bayes , logistic regression and gradient boosting classifier respectively.Below are the results achieved with these in order :



NAIVE BAYES



LOGISTIC REGRESSION



GRADIENT BOOSTING CLASSIFIER

Ultimately, we assess the model performance by employing a confusion matrix, and subsequently utilize it to compute essential metrics such as precision, recall, and F1 score, providing a comprehensive evaluation to determine the superior model.

Confusion Matrix : A confusion matrix is a tabular representation used in machine learning to evaluate the performance of a classification model. It summarizes the counts of true positive, true negative, false positive, and false negative predictions, providing insights into the model's accuracy and errors.

Precision, recall, and F1 score are metrics derived from a confusion matrix in AI:
Precision: Precision is the ratio of true positive predictions to the total predicted positives, indicating the accuracy of positive predictions.It is seen as true positive divided by sum of true positive and false positive
Recall (Sensitivity): Recall measures the ability of a model to capture all the actual positive instances. It is the ratio of true positive predictions to the total actual positives and is calculated true positive divided by sum of true positive and false negative
F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balanced measure between precision and recall.
Accuracy: Accuracy measures the overall correctness of predictions and is the ratio of correct predictions (true positives and true negatives) to the total predictions.

Chapter 5: Results and Conclusions

After extensive experimentation, the sentiment analysis project yielded noteworthy outcomes. Notably, the Gradient Boosting Classifier exhibited the highest overall precision among the models assessed. Precision, recall, and F1 score were computed using the confusion matrix for each model.

The observed dominance of Gradient Boosting in precision suggests its adeptness at correctly classifying positive and negative sentiments. This can be attributed to the model's ability to capture intricate relationships within the data. Additionally, the wordcloud visualizations provided insightful representations of sentiment patterns.

The preprocessing steps, including punctuation removal, stemming, and stopword exclusion, played a pivotal role in enhancing the dataset's cleanliness. Count vectorization further facilitated the transformation of textual data into a numerical format suitable for machine learning models.

While each model—Naive Bayes, Logistic Regression, and Gradient Boosting—demonstrated competitive performance, the nuanced strength of Gradient Boosting in precision underscores its efficacy in discerning sentiment complexities within the Twitter dataset.

Conclusion:

In conclusion, this sentiment analysis project successfully employed advanced techniques to decipher sentiments expressed in Twitter data. The thorough evaluation revealed Gradient Boosting as the model with the highest precision, emphasizing its proficiency in accurately categorizing sentiments.

The significance of this project lies in its contribution to refining sentiment analysis methodologies for Twitter, providing valuable insights for diverse applications. As social media continues to evolve, the ability to precisely interpret sentiments becomes increasingly crucial for businesses, policymakers, and researchers.

These findings pave the way for future research aimed at enhancing the interpretability and generalization of sentiment analysis models, ensuring their effectiveness across various domains.

REFERENCES

[1]Dey, L., Chakraborty, S., Biswas, A., Bose, B. and Tiwari, S., 2016. Sentiment analysis of review datasets using naive bayes and k-nn classifier. *arXiv preprint arXiv:1610.09982*.

[2]Saleena, Nabizath. "An ensemble classification system for twitter sentiment analysis." *Procedia computer science* 132 (2018): 937-946.

[3] Ramadhan, W.P., Novianty, S.A. and Setianingsih, S.C., 2017, September. Sentiment analysis using multinomial logistic regression. In *2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC)* (pp. 46-49). IEEE.

[4]Hew, K.F., Hu, X., Qiao, C. and Tang, Y., 2020. What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education*, *145*, p.103724.