صورة تحتوي على نص, شعار, لقطة شاشة, التصميم

تم إنشاء الوصف تلقائياً



**Task 1: Text Analysis**

**"Twitter Airline Sentiment Dataset"**

**DS3114))**

**Dr. Omaima Fallatah))**

|  |  |
| --- | --- |
| **ID:** | **Name:** |
| **444005403** | **Sara Ayed Alsehli** |
| **443009106** | **Lina Ibrahim Alghamdi** |

**Table of Contents:**

[Introduction 3](#_Toc180839867)

[Dataset Overview: 3](#_Toc180839868)

[Data Preprocessing 5](#_Toc180839869)

[Lowercasing: 5](#_Toc180839870)

[Punctuation and Special Characters Removal: 5](#_Toc180839872)

[Tokenization: 5](#_Toc180839874)

[Stopword Removal: 5](#_Toc180839876)

[Stemming: 5](#_Toc180839878)

[TF-IDF Vectorization: 5](#_Toc180839880)

[Model Performance 6](#_Toc180839882)

5 [Naive Bayes(Multinomial):](#_Toc180839883)

5 [Random Forest:](#_Toc180839884)

6 [Logistic Regression:](#_Toc180839885)

6 [SVM (Support Vector Machine):](#_Toc180839886)

6 [XGBoost:](#_Toc180839887)

[Insights 8](#_Toc180839888)

[Imbalance Impact: 8](#_Toc180839889)

[Model Suitability: 8](#_Toc180839890)

[Importance of Text Features: 9](#_Toc180839891)

[Conclusion 10](#_Toc180839892)

[References 11](#_Toc180839893)

Introduction

Introduction This task involves analyzing the Twitter Airline Sentiment dataset, which consists of 14,640 tweets categorized into three sentiment labels: positive, neutral, and negative. The dataset contains columns such as "tweet\_id, airline\_sentiment, airline\_sentiment\_confidence", and the actual text of the tweets including information about tweet content, sentiment, and other metadata. The task involves text preprocessing, model training, and evaluating model performance. The objective is to preprocess the text data and apply machine learning models to predict the sentiment of tweets.

## **Dataset Overview:**

The Twitter Airline Sentiment dataset contains **14,640 rows** and **15 columns**. Key columns include:

* **tweet\_id:** A unique identifier for each tweet.
* **airline\_sentiment:** The sentiment of the tweet (positive, negative, or neutral).
* **airline\_sentiment\_confidence:** Confidence level of the sentiment classification.
* **negativereason:** The reason for negative sentiment (if applicable).
* **airline:** The airline company mentioned in the tweet.
* **retweet\_count:** Number of retweets.
* **text:** The actual content of the tweet.

The dataset focuses on analyzing user sentiment towards airlines based on tweets, aiming to classify the sentiment and identify reasons behind negative feedback.

# **Data Preprocessing**

To prepare the data for modeling, several preprocessing steps were applied to the tweet text:

## **Lowercasing:**

## Converted all text to lowercase to maintain consistency.

## **Punctuation and Special Characters Removal:**

## Removed non-alphabetic characters, URLs, and extra space.

## **Tokenization:**

## Split the text into individual words.

## **Stopword Removal:**

## Removed common English stopwords that do not contribute to sentiment.

## **Stemming:**

## Reduced words to their root form using (PorterStemmer).

## **TF-IDF Vectorization:**

## Converted the processed text into numerical features using **TF-IDF (Term Frequency-Inverse Document Frequency)** to capture important words in the context of the corpus.

# 

# **Model Performance**

# 

Several models were tested to predict tweet sentiment:

* **Naive Bayes (Multinomial):**

**Accuracy**: 73.4%

**Strength:** Effective for simple text classification.

**Weakness:** Lower performance for minority classes like neutral and positive sentiments.

* **Random Forest:**

**Accuracy:** 75.5%

**Strength:** Better at handling class imbalance.

**Weakness:** Computationally more intensive.

* **Logistic Regression:**

**Accuracy:** 78.1%

**Strength:** Balanced performance across all classes.

**Weakness:** Slightly lower precision for minority classes.

* **SVM (Support Vector Machine):**

**Accuracy:** 77.6%

**Strength:** High precision for negative sentiment.

**Weakness:** Lower recall for neutral and positive sentiments.

* **XGBoost:**

**Accuracy:** 73.9%

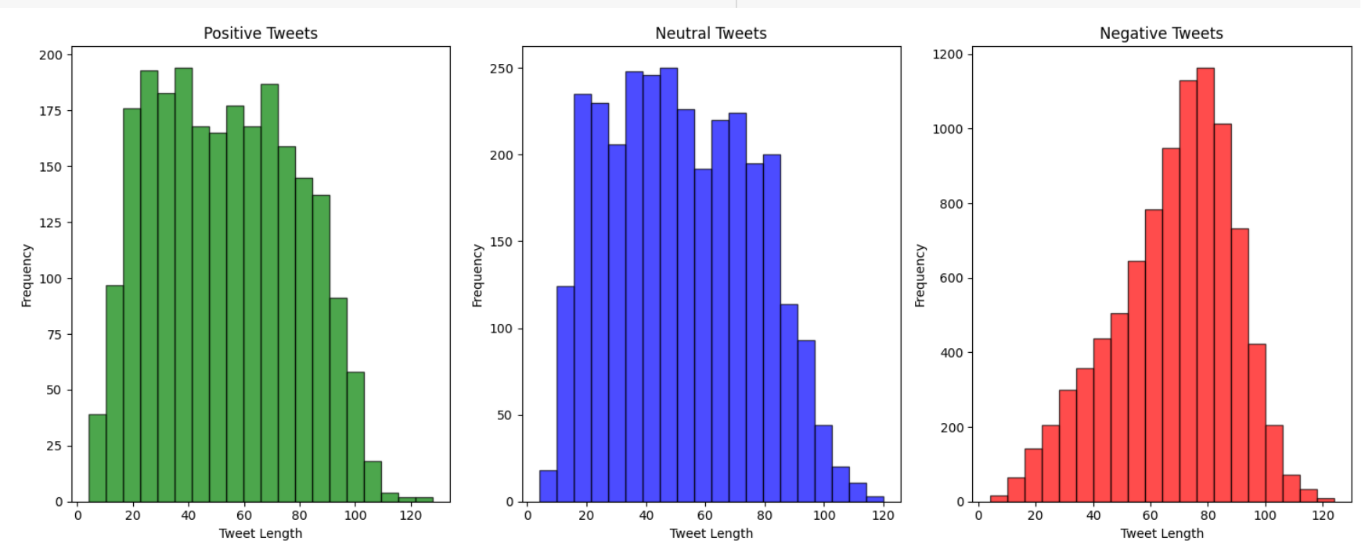
**Strength:** Good handling of complex decision boundaries.

**Weakness:** Slower training time compared to other models.

**The best-performing model** in this analysis was **( Logistic Regression)**, achieving an accuracy of **78.1%.** This success is due to the model’s ability to effectively distinguish between different sentiment categories using the logistic function, which enhances classification of high-dimensional text data, especially when transformed using TF-IDF. This method leverages linear separation between classes, significantly improving the model's accuracy in effectively classifying tweets.

# **Insights**

## **Imbalance Impact:**

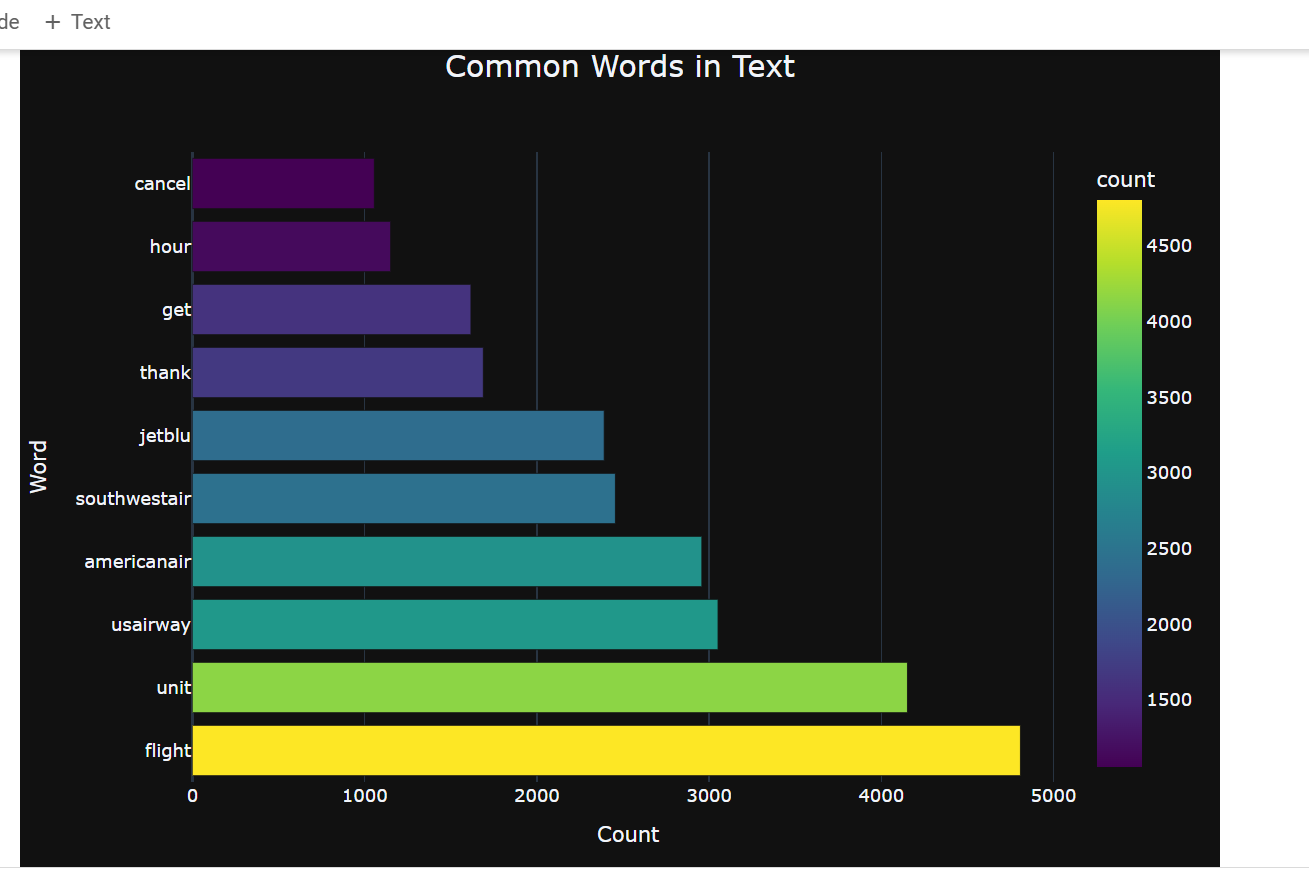
 The dataset's imbalance, with a high proportion of negative tweets, impacted model accuracy for positive and neutral sentiments. Techniques like SMOTE helped balance class representation but require additional tuning for optimal results.

## **Model Suitability:**

Logistic Regression and SVM performed well in this context, effectively managing the high-dimensional text data from TF-IDF transformation, which helped improve prediction accuracy across different sentiment categories. While Naive Bayes struggled with neutral and positive classes.

## **Importance of Text Features:**

The use of TF-IDF highlighted key words and phrases, enhancing the model's ability to differentiate sentiments by focusing on contextually significant terms.



# 

# **Conclusion**

This analysis shows that Logistic Regression and SVM provided the most balanced performance. Handling the class imbalance is crucial for improving model accuracy and recall across all sentiment categories. While effective, but it is possible that further improvements can be made by addressing class imbalance and exploring deeper models like neural networks. Overall, the study lays strong groundwork for future sentiment analysis on social media data.

# **References**

(<https://www.kaggle.com/crowdflower/twitter-airline-sentiment>)