Sentiment Analysis of Airline Tweets

Abstract

The project performs sentiment analysis on the tweets mentioning each major U.S. airline. Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as “late flight” or “rude service”). In our project we used our model, which we trained using a naïve bayes classifier to predict if the tweets were either positive, negative or neutral.
We compare our predictions with the user-classified versions of those tweets, and based on this information, we produce a classification report that includes F1 scores, precision, accuracy, and other metrics. The user has the option to input some tweets and evaluate the model after it has finished running. The train dataset and test dataset are randomized during each execution which doesn’t decrease the probability of the code.
For the Phase-1 We are focused on Preprocessing data, Performing EDA and Extracting Significant Insights based on data.

Introduction

Opinion or assessment motivated by emotion is called sentiment. Sentiment analysis involves extracting the opinion of the user on a scale based on the feedback given by the user. Any kind of event or category can be subjected to sentiment analysis, and the results can be used to improve the situation to a desirable degree.

We used Twitter data to conduct sentiment analysis as part of our project. Major airline Twitter users were the source of this data. The airlines can enhance those characteristics by extracting the keywords from the bad tweets. For example, airlines can concentrate on correctly scheduling them and ensuring that they arrive on time if the main complaint expressed in the bad tweets is delays. This aids in

highlighting the airline company's advantages but can also be utilized to find the tweets that are positive.

Objective

Objective is to identify the sentiment of tweets generated by the users on Twitter. Our program would train a model which would then perform sentiment analysis on the above-mentioned data. It would be able to identify the sentiment with as much accuracy as possible.

Motivation

Many times, on twitter we see many trending hashtags, but in order to understand whether they are trending, either for negative or positive reasons is hard. The tweets which pop up soon after we click the tag might be negative but not all of them are the same. Finding majority opinion is hard. Given the data frame our program will be able to identify the sentiment and it would be easy to identify why the tag is trending.

We can use the code for various data frames, it’s not limited to airline datasets.

Contribution value

The contribution value of this project is its ability to provide customer feedback to the airlines with a data-driven method in real time. By using sentiment analysis, airlines can understand the mood of their customers and respond immediately to any kind of negative comments they receive. It helps in identifying the trends, common issues faced, etc. which can be helpful in improving the service delivery.

Customer feedback: This project gives airlines a medium through which they can know the emotions and feedback patterns of the customers and is essential to improve the brand value and customer experience.
Improvements in operations: The information drawn from customers can be used to coincide with the operational data to identify the common mistakes, failures and make better resource allocation and solving problems.

Brand value: This analysis helps airlines understand the current issues, trends quickly, enabling them to respond to the negativity spreads all over. This helps in protecting the brand image.

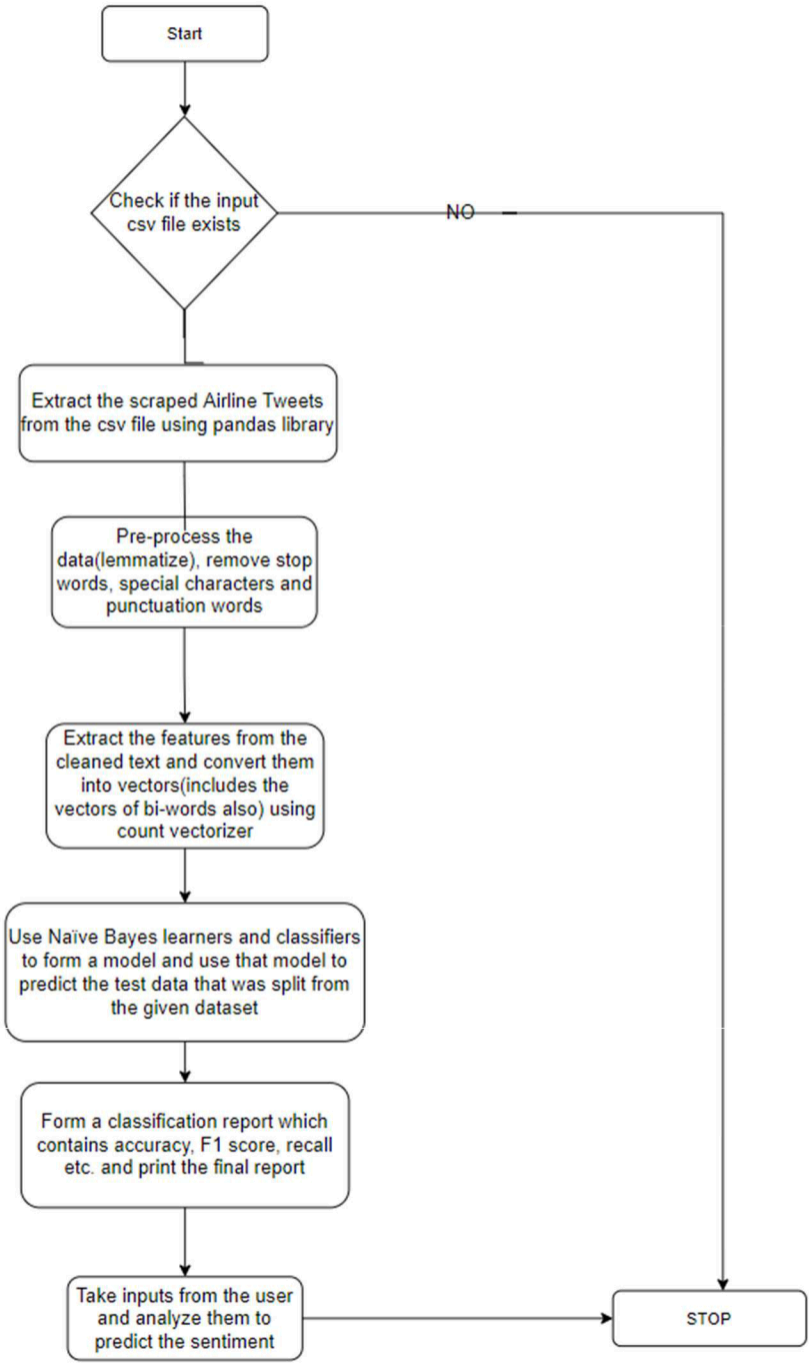
Why does it warrant investigation?

Customer service relevant: Airlines business is a highly competitive business where customer service can be a main factor that decides the brand image. Therefore, understanding how the customers react in real time and adapting their services according to that will be very crucial for airlines.

Real time feedback: As we are using twitter, which is the most used platform where people express immediate thoughts and expression in real time, this makes it a rich source for real time feedback.
Impact of Social Media: Social media has become a significant part in creating a certain perception in

people about anything. This allows airlines to investigate on the customer sentiments ad respond accordingly and improve their customer interaction,
Crisis management: few cases like, during delays, cancellations, or poor customer service, the tweets are a direct indication of the people's opinion. This makes sentiment analysis a required tool for crisis management in airlines.

Proposed Model



Methodology

Data Extraction Methodology

1. Source of Data :

- **Kaggle Dataset:** The data for this project was obtained from Kaggle, a popular platform for datasets and data science projects.
- **Dataset Details:**
    - The dataset is titled "**Twitter US Airline Sentiment.**"
    - It contains tweets from February 2015, mentioning major U.S. airlines.
    - Includes pre-labeled sentiment categories: **positive, neutral** and **negative.**

2. Data Acquisition Process :

1. **Downloading the Dataset:**
    - The dataset was downloaded directly from Kaggle.
    - **File format:** CSV, which is easy to load and manipulate using data analysis tools like Pandas.
2. **Dataset Description:**
    - **Columns:**
        - Key columns include:
            - **text:** The tweet content.
            - **airline_sentiment:** Sentiment classification of tweets.
            - **Negative reason:** Reasons for negative sentiment, if applicable.
            - Metadata such as **tweet_created**, **airline**, and **user_timezone**.
    - Data Size: 14,640 tweets, covering various sentiment types.

3. Advantages of Using Kaggle Data

- **Pre-Cleaned and Pre-Labeled**:
    - The dataset is curated, with tweets already categorized as positive, neutral, or negative.
    - Negative tweets are further categorized by specific reasons.
- **Reliable Source**:
    - Kaggle datasets are well-documented, making them suitable for academic and professional projects.

4. Challenges and Steps Taken

1. **Relevance**:
    - Ensuring the dataset aligns with the project objectives.
    - Focused on extracting insights specific to airline sentiment.
2. **Preprocessing**:
    - Additional cleaning was performed to refine the dataset further.
    - Text cleaning steps included removing duplicates, punctuation, and irrelevant characters.

Data cleaning

Removing Stop words:

Importance: Stop words (common words like "the", "is", "in") are often irrelevant in determining the sentiment of a tweet. Removing them helps the model focus on more meaningful words that contribute to the overall sentiment.

Removing URLs:

Importance: URLs usually do not provide any meaningful insight into the sentiment of the text and might distort the model's understanding. Removing them ensures cleaner text data for processing.

Removing Punctuation:

Importance: Punctuation marks like periods, commas, and exclamation marks are generally not informative for sentiment classification. Removing them makes the text easier to process and analyze.

Removing HTML Tags:

Importance: Tweets or text data might contain HTML tags when retweeting or quoting other content. Removing these tags ensures the model only focuses on actual text content.

Removing Twitter Usernames (@username):

Importance: Mentions of usernames are common in tweets but do not convey meaningful information about the sentiment. Removing them helps in preventing unnecessary noise.

Removing Emojis:

Importance: While emojis can express sentiment, they may introduce complexity during text processing. Removing them simplifies the text and avoids confusion for the model.

Text Abbreviation Expansion:

Importance: Twitter language often includes abbreviations and contractions (e.g., "can't", "won't").Expanding these to their full forms ensures that the model interprets the text accurately.

Removing Numbers:

Importance: Numbers generally don’t provide sentiment insight, so removing them ensures cleaner data for text processing.

Handling Continuously Repeated Characters:

Importance: Repeated characters (like "hiiiiii") are common in informal texts. Normalizing them ensures better text representation for sentiment analysis

Removing Non-Alphabetical Characters:

Importance: Removes any characters that are not alphabetic, simplifying the text for analysis.

Combining Negative Reason with Tweet:

Importance: If a tweet is tagged with a negative reason, combining this with the text ensures that the reason is considered while analyzing sentiment.

EDA insights

1. Sentiment Distribution:

The dataset contains 14,640 tweets categorized into three sentiment classes:

- **Negative**: 9,178 tweets
- **Neutral**: 3,099 tweets
- **Positive**: 2,363 tweets

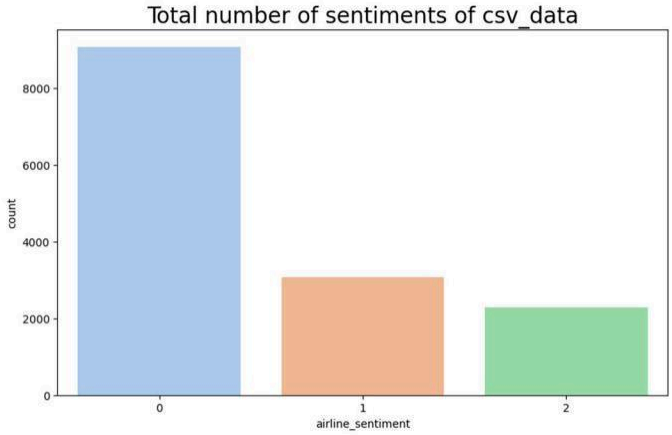
The dataset exhibits a significant class imbalance, with **negative tweets** dominating. This imbalance could lead to biased performance of machine learning models, necessitating the use of techniques such as **SMOTE** (Synthetic Minority Oversampling Technique) to balance the classes.

What is SMOTE?

- **SMOTE (Synthetic Minority Oversampling Technique)** is a data augmentation method used to handle class imbalance.
- Instead of simply duplicating examples from the minority classes, SMOTE generates **synthetic examples** by:
    - Identifying the nearest neighbors of a sample in the feature space.
    - Creating new samples as linear combinations of the original and its neighbors.
- This results in a **balanced dataset** that maintains the diversity of the original data.

Why We Used SMOTE?

- **Preserves Information from the Dataset**:
    - Unlike random oversampling, which duplicates data and may cause overfitting, SMOTE generates diverse synthetic samples, preserving the overall structure of the dataset.
- **Improves Model Training**:
    - A balanced dataset ensures that the model receives equal attention to all classes, improving the ability to classify neutral and positive sentiments.
- **Enhances Metric Performance**:
    - SMOTE reduces bias in evaluation metrics such as accuracy, precision, recall, and F1-score by addressing the dominance of the negative sentiment class.

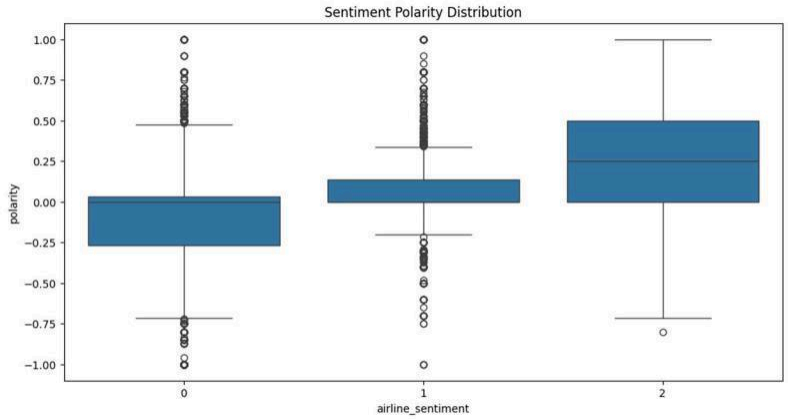


2. Sentiment Polarity:

The polarity scores computed using the TextBlob library are well aligned with the sentiment labels given manually.

Therefore, polarity is likely to serve as a useful feature for the model.

It was observed that tweet length correlates with sentiment polarity, suggesting that longer tweets may express different sentiments compared to shorter ones.

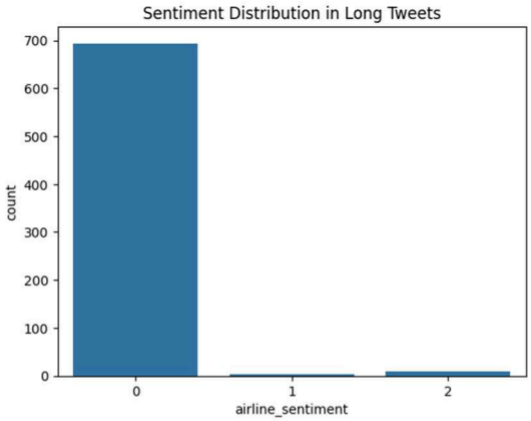


3. Tweet Length Analysis:

The average tweet is about 67 characters long, with a standard deviation of 30.24. Tweet lengths range from a minimum of 0 to a maximum of 155 characters.

The distribution of tweet lengths is unimodal and roughly symmetric, with a skewness of -0.23, indicating that there is roughly the same spread above and below the mean.

The interquartile range (IQR) for tweet length is 47, reflecting the spread between the 25th percentile (44 characters) and the 75th percentile (91 characters), There are no significant outliers in tweet lengths, indicating that most tweets fall within a reasonable length range.



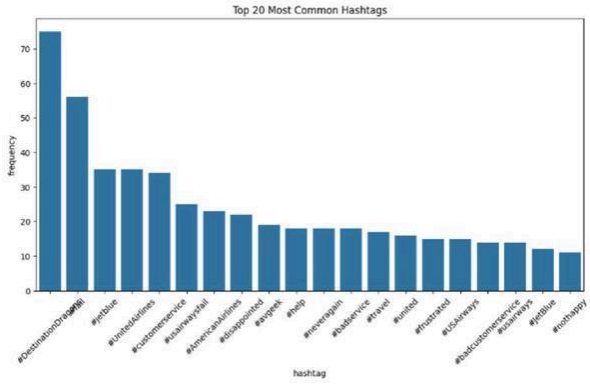

```
Number of outliers in tweet length: 0
Percentage of outliers: 0.00%
```

4. Hashtag Analysis:

Hashtag analysis conducted on the dataset revealed common themes or topics that could be useful in feature engineering. Hashtags typically indicate the core sentiment or topic of a tweet, making them valuable for sentiment classification.

The top 20 most frequent hashtags highlight trending topics and concerns among users, such as airline service quality, delayed flights, or expressions of appreciation.

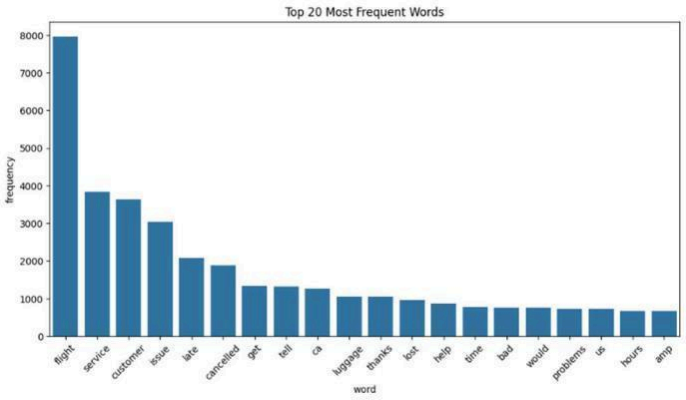
Common hashtags are: fail,help, not happy and few airline names suggesting that many airline users also use twitter to post concerns regarding their travel.



5. Word Frequency Analysis:

Word frequency analysis provides common terms and phrases that can be used to create features, helping the model understand the meaning of tweets.

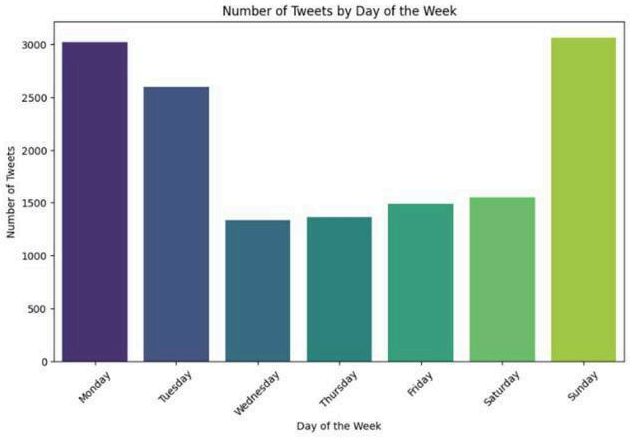
Frequently occurring words like "delayed" or "thank" are likely references to common complaints or expressions of gratitude, which are important for differentiating sentiments.



6. Time-Based Analysis:

Analysis of tweet activity over the week shows that tweet counts vary across different days, peaking on some days. This could be valuable for time-based feature engineering to capture time-dependent sentiment.

Time-of-day trends might indicate events or announcements from airlines, offering additional context for understanding shifts in sentiment.

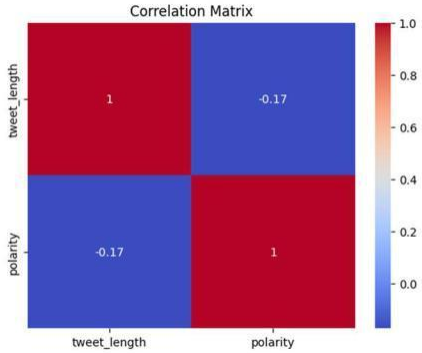


This suggests that Travelling peaks during the start and end dates of the week.

7. Correlation Analysis:

The correlation matrix between tweet length and polarity shows a very weak relationship, suggesting that while tweet length may not directly indicate polarity, it can provide some insights into the nature of sentiment.

Further correlations, such as those between the day of the week and sentiment, could reveal additional relationships that might enhance the model.



8. Outlier Analysis:

There are no significant outliers in tweet lengths, indicating that tweet lengths remain relatively consistent across the dataset.

The absence of outliers simplifies the data preprocessing process, as no special handling is required for tweet lengths.

9. Distribution Skewness:

The distribution of tweet lengths is near-symmetric, with minimal skewness. This means that most tweets are of similar length, reducing the need for transformations such as log scaling.

A symmetric distribution simplifies modeling since many algorithms assume this type of distribution, improving the model's performance.

```
Number of peaks in tweet length distribution: 1
The distribution appears to be unimodal.

Skewness of tweet length distribution: -0.23
The distribution is approximately symmetric.
```

10. Multimodality of Data:

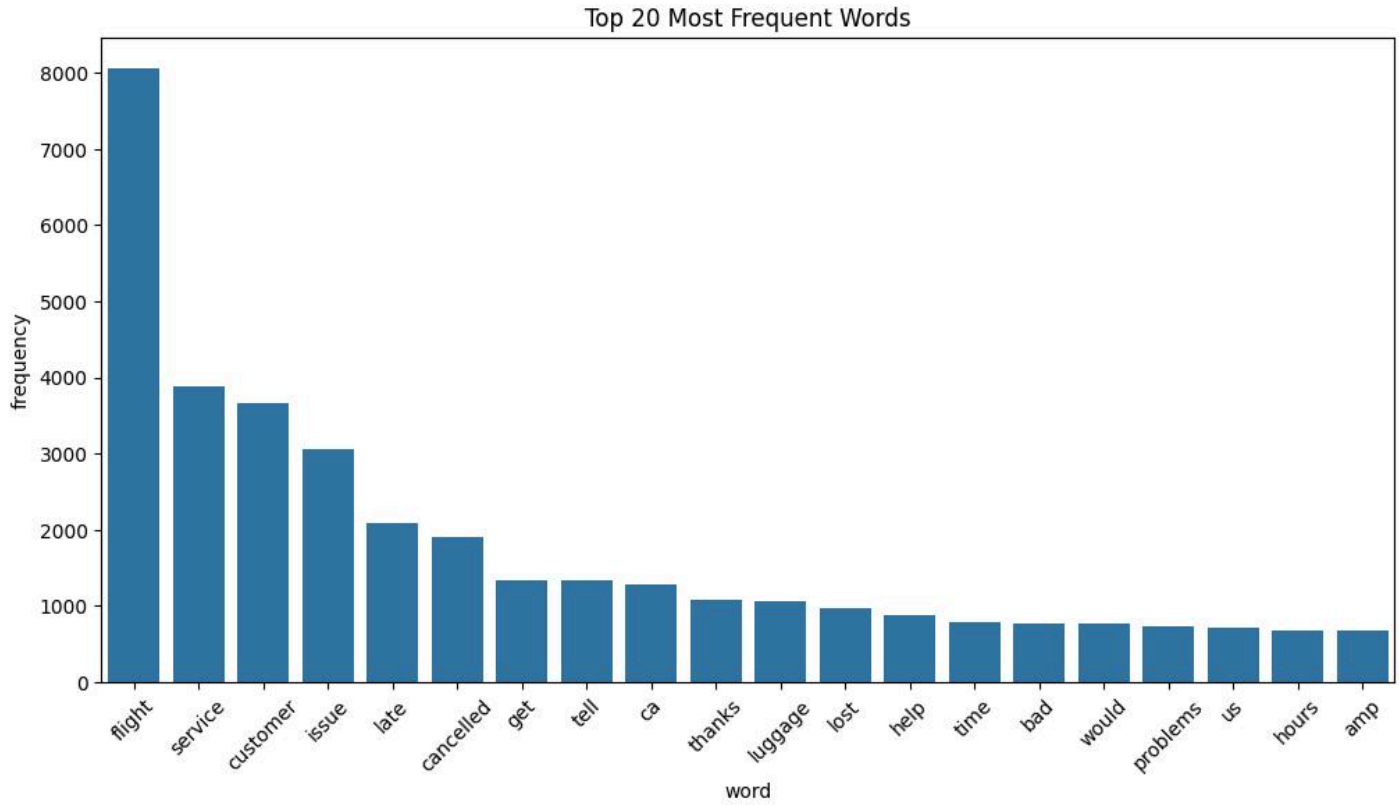
The analysis shows that tweet lengths follow a unimodal distribution rather than a bimodal distribution. Users tend to tweet at similar lengths rather than clustering into separate length groups.

Understanding this behavior provides insights into fine-tuning features based on typical tweet lengths, which may help improve the accuracy of sentiment classification.

Word Frequency Analysis :

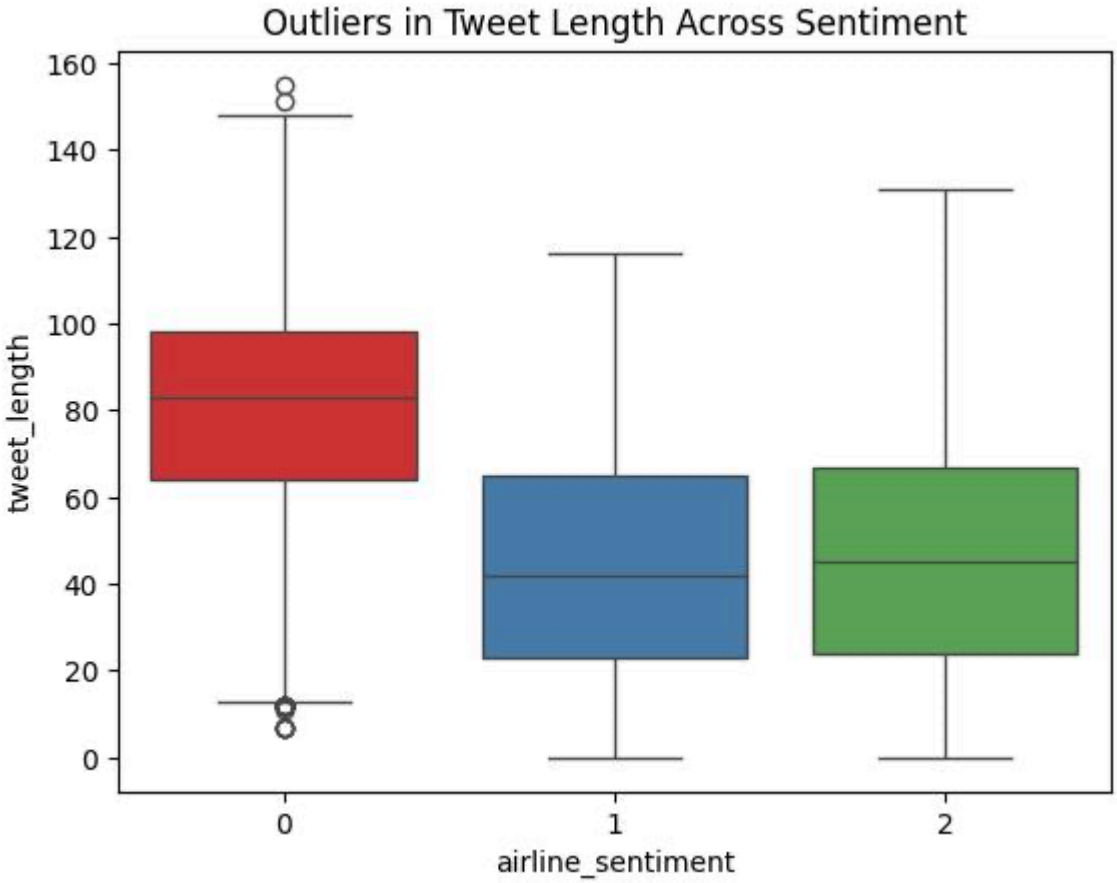
The dataset underwent a comprehensive word frequency analysis to uncover the most frequently occurring terms. By analyzing the corpus of tweets, it was observed that words such as "flight," "thank," "delay," "service," and "cancelled" appeared prominently. These high-frequency words correspond to the core themes of customer interaction, including both complaints and expressions of gratitude.

To visualize this, a bar chart displaying the top 20 most common words was created, highlighting the prevalence of terms that reflect operational issues or customer appreciation.

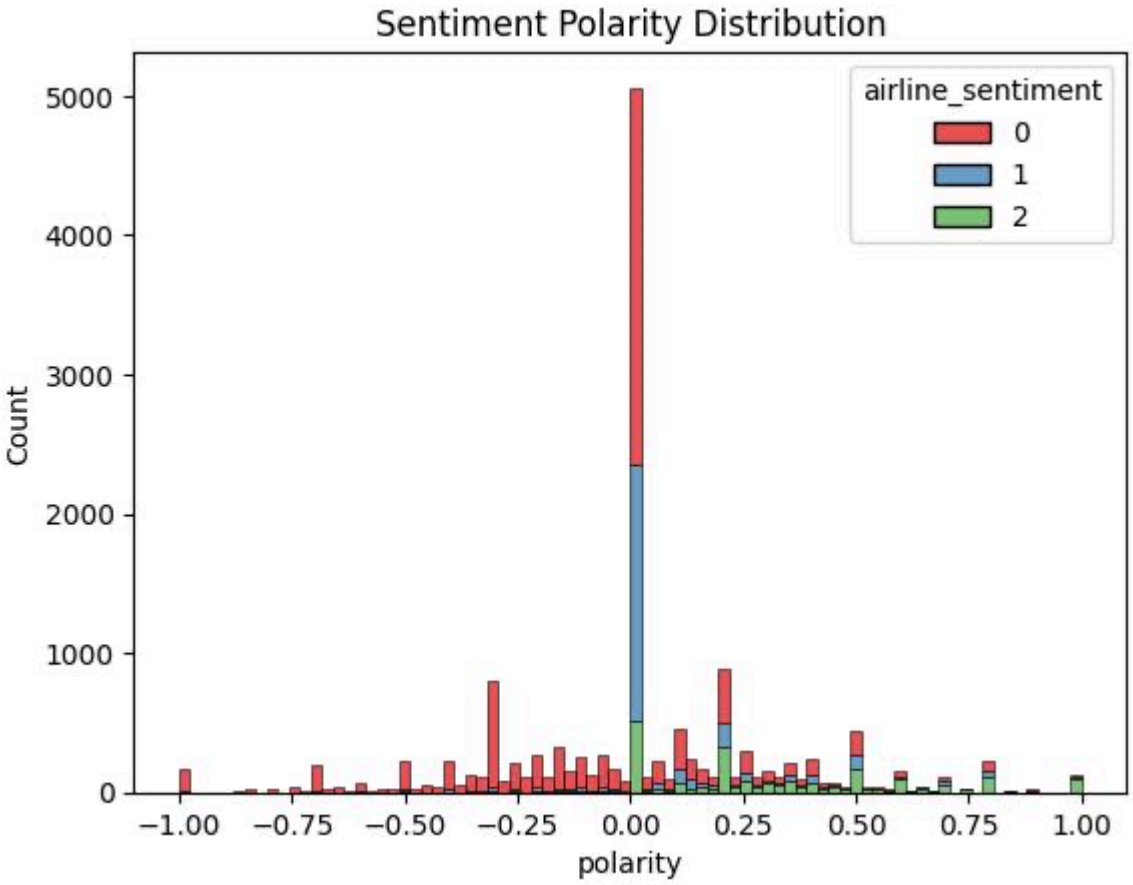


Feature scaling and selection were performed to prepare the data for machine learning. The features, including tweet length and polarity, were standardized using a scaler to eliminate discrepancies in scale. Mutual information was applied to identify the most informative features for predicting sentiments. This process confirmed that tweet length and polarity are key indicators, significantly influencing the sentiment classification model.

Further analysis focused on tweet length revealed outliers, particularly in the negative sentiment class, where tweets exhibited considerable variability in length. Positive sentiments were more balanced, while neutral sentiments displayed minimal variation. A box plot was used to illustrate these differences, underscoring the importance of long tweets in capturing nuanced sentiments, whether positive or negative.



The sentiment polarity distribution, categorized by sentiment classes, was also explored. A histogram of polarity values demonstrated a clear alignment between the calculated polarity and manually labeled sentiments.



Positive sentiments clustered around higher polarity values, negative sentiments around lower polarity values, and neutral sentiments near zero. This distribution validates the use of polarity as a robust feature for sentiment classification.

These analyses collectively enhance the understanding of the dataset's structure and inform the feature engineering process. The insights derived, particularly the importance of tweet length, polarity, and common terms, provide a strong foundation for building an effective sentiment classification model.

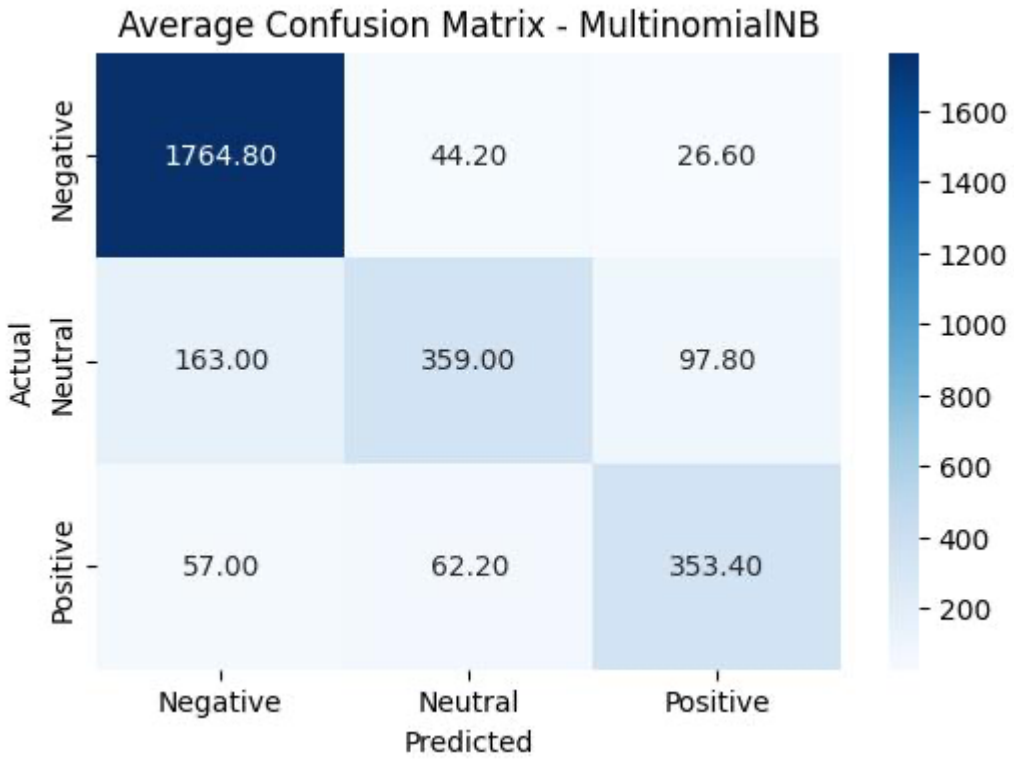
Model Performance and Insights :

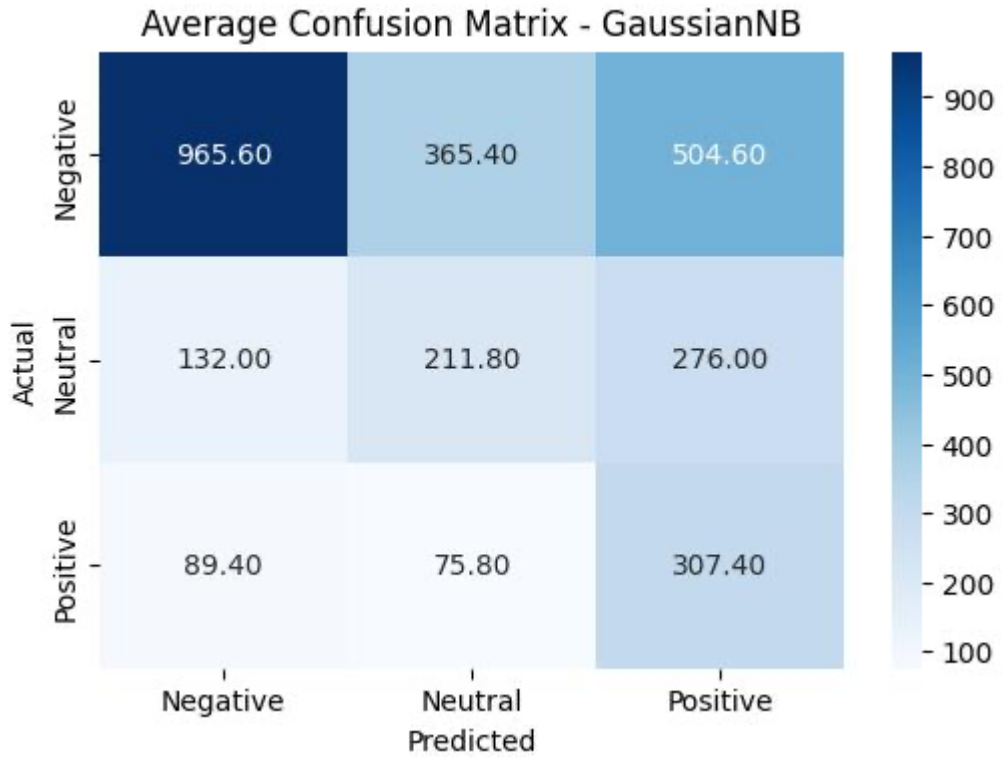
Three classification models—**Multinomial Naïve Bayes (MultinomialNB)**, **Gaussian Naïve Bayes (GaussianNB)**, and **K-Nearest Neighbors (KNN)**—were evaluated using stratified K-Fold cross-validation with SMOTE to address the class imbalance. The models were assessed on their ability to classify tweets into three sentiment classes: negative, neutral, and positive.

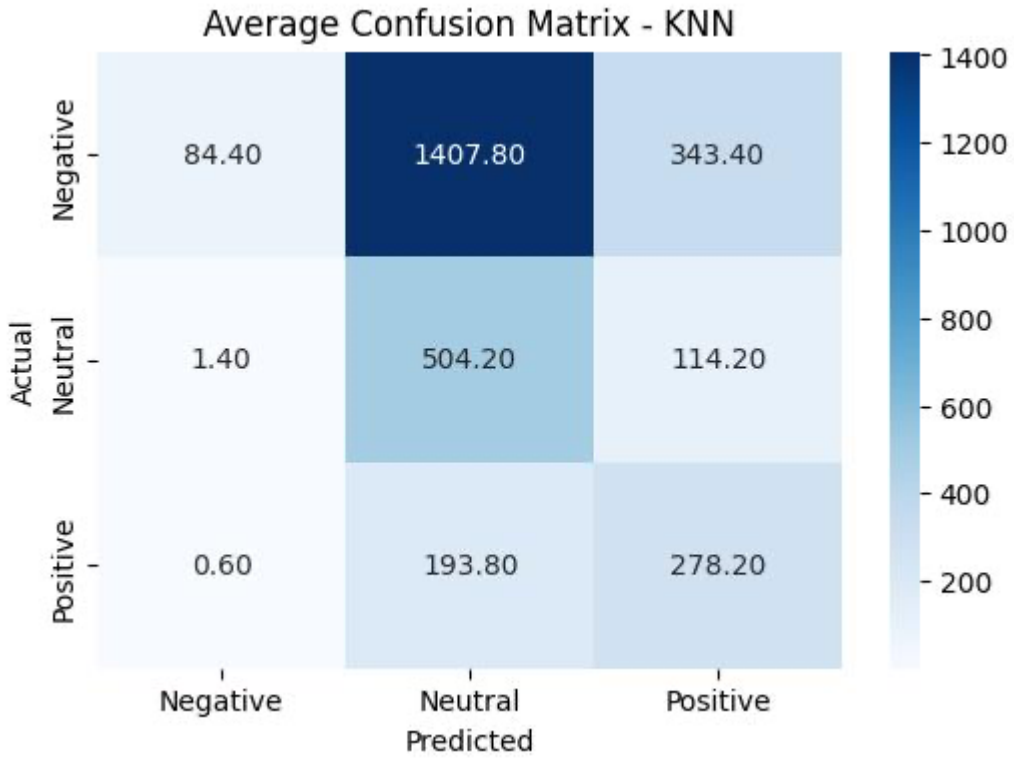
Model	Accuracy	Precision	Recall	F1-Score
MultinomialNB	84.60%	80.01%	76.28%	77.63%
GaussianNB	50.71%	47.36%	50.61%	45.52%
KNN	29.60%	54.10%	48.27%	30.78%

The results demonstrated that MultinomialNB achieved the highest performance across all metrics. Its average accuracy was **84.60%**, with an F1-score of **77.63%**, outperforming GaussianNB and KNN significantly.

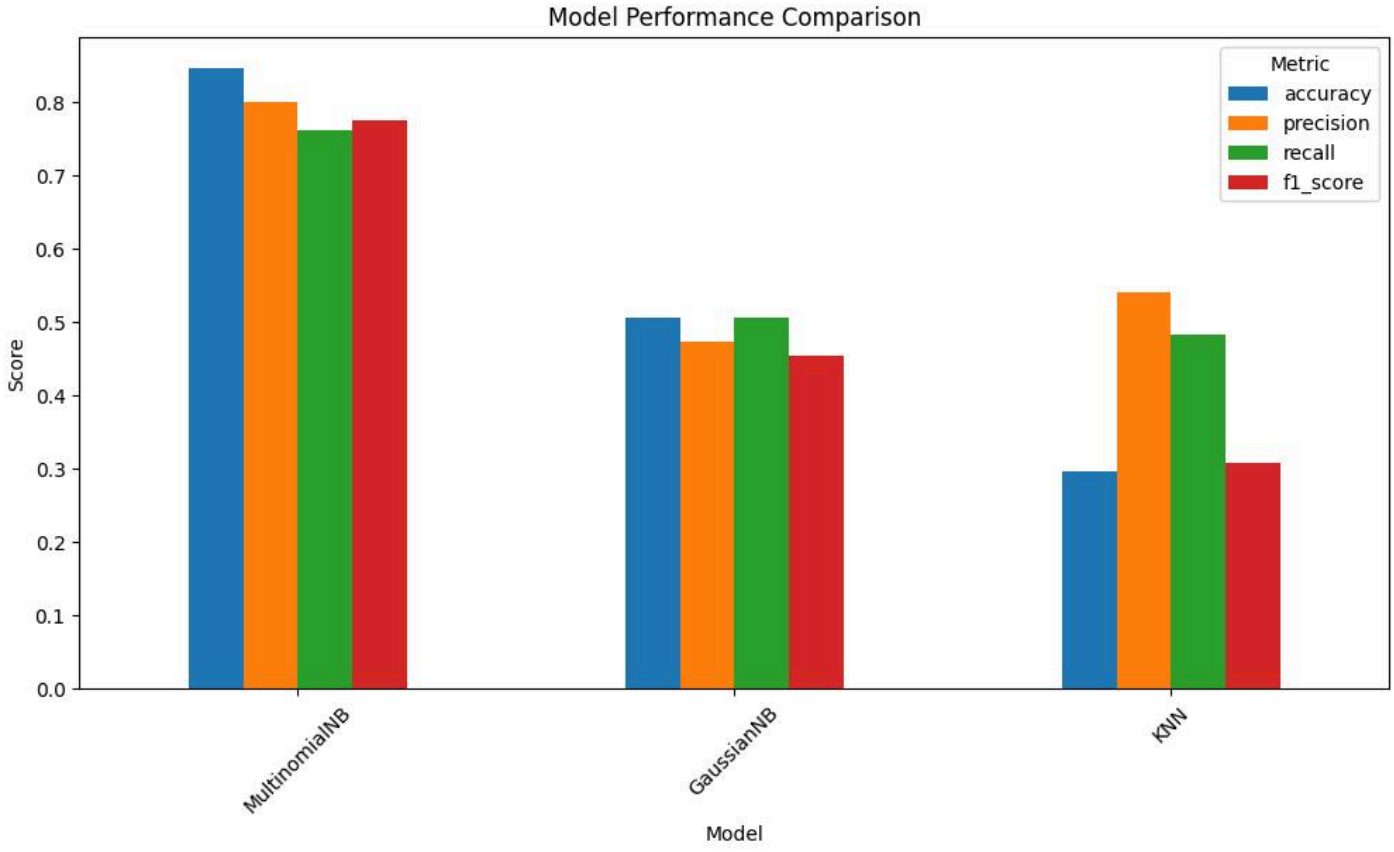
The confusion matrix for MultinomialNB indicated strong classification performance for negative sentiments, with limited misclassification of neutral and positive tweets. GaussianNB, while performing better than KNN, showed difficulty in distinguishing between neutral and positive sentiments. KNN struggled to handle the high-dimensional TF-IDF feature space, resulting in the lowest accuracy of **29.60%** and inconsistent performance across folds.



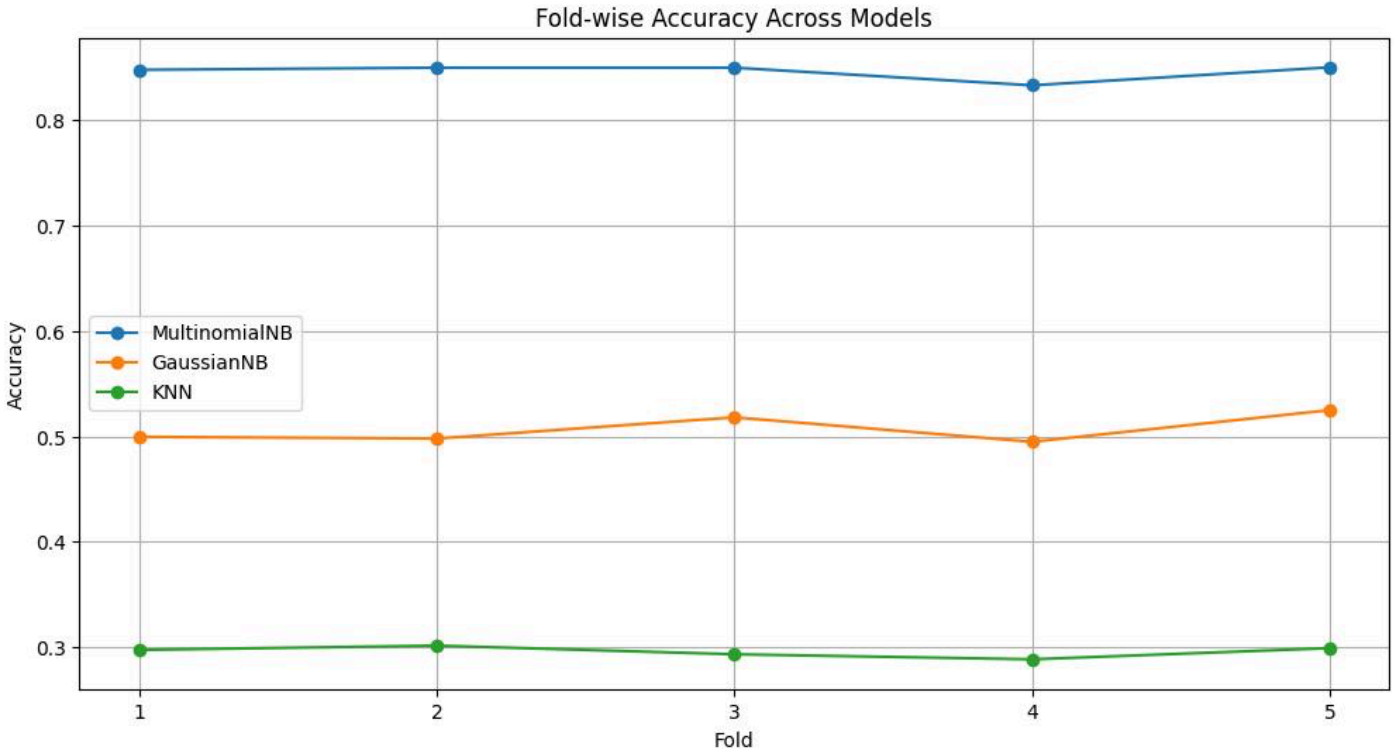

Average Confusion Matrix - GaussianNB
Actual
Negative
Neutral
Positive
965.60
365.40
504.60
132.00
211.80
276.00
89.40
75.80
307.40
900
800
700
600
500
400
300
200
100
Negative
Neutral
Positive
Predicted


Average Confusion Matrix - KNN
Actual
Negative
Neutral
Positive
84.40
1407.80
343.40
1.40
504.20
114.20
0.60
193.80
278.20
1400
1200
1000
800
600
400
200
Negative
Neutral
Positive
Predicted

A comparison of average confusion matrices provided further insights. MultinomialNB had the highest true positive rates across all sentiment classes, particularly for negative sentiments, which accounted for the majority of the dataset. GaussianNB exhibited higher misclassification rates, often confusing neutral tweets as positive. KNN, due to its sensitivity to high-dimensional data, misclassified a significant portion of both neutral and positive sentiments.



The bar chart comparing metrics further highlighted the superiority of MultinomialNB, with consistent performance across all evaluation criteria, including precision, recall, and F1-score. The fold-wise accuracy plot reinforced this consistency, showing minimal variance in MultinomialNB's performance across all five folds. Conversely, GaussianNB displayed moderate variability, while KNN exhibited the most unstable results.



These findings affirm that MultinomialNB is the most suitable model for sentiment classification in this context. Its ability to effectively handle sparse, high-dimensional TF-IDF features ensures robust predictions across all sentiment classes. Future enhancements could involve the integration of advanced ensemble techniques, such as Gradient Boosting or Random Forest, to explore potential performance gains. Additional features, such as temporal trends and hashtags, could also be incorporated to enrich the feature set and further refine classification accuracy.

The finalized Multinomial Naïve Bayes (MultinomialNB) model, identified as the best-performing sentiment analysis classifier based on its F1-score, was deployed for real-time testing and prediction. The deployment includes an interactive user interface that allows users to input custom text and receive predictions for both sentiment and tone (sarcasm).

Real-Time Sentiment Prediction Workflow :

1. **Selection of the Best Model:**
   - The model with the highest F1-score, **MultinomialNB**, was chosen for deployment. Its superior performance across all evaluation metrics ensures reliable predictions.
2. **Input Preprocessing:**
   - User input is preprocessed using the same techniques applied during training to ensure consistency:
     - **Stopword Removal**: Eliminates irrelevant words.
     - **URL, Punctuation, and HTML Tag Removal**: Cleans text for better analysis.
     - **Username and Emoji Removal**: Strips mentions and emotive symbols.
     - **Text Abbreviation Expansion**: Converts abbreviations like "can't" to "cannot."
     - **Number Removal**: Excludes numeric values to focus on text content.
3. **Feature Vectorization:**
   - The preprocessed text is transformed into numerical feature vectors using the **TF-IDF vectorizer**, consistent with the training data preparation.
4. **Sentiment Prediction:**
   - The MultinomialNB model predicts the sentiment of the user-provided text as one of three classes:
     - **Negative** (0), **Neutral** (1), or **Positive** (2).
   - If the model supports probability outputs (e.g., predict_proba), the prediction probabilities are provided for greater transparency.
5. **Sentiment Verdict:**
   - The sentiment class with the highest probability is highlighted as the **final sentiment verdict**, ensuring interpretability.
6. **Tone Detection:**
   - In addition to sentiment prediction, a separate sarcasm detection model is applied. This allows for more nuanced tone analysis by identifying potentially sarcastic inputs.

User Interaction :

The system allows users to interactively test the sentiment model:

- Users input text for analysis.
- The system preprocesses the input and provides:

    - The sentiment prediction and probabilities.
    - A tone (sarcasm) detection verdict.

The system also includes a termination mechanism, where users can exit by typing "quit."

Example Predictions:

- **Input**: "The flight was delayed, but the crew was very polite."
    - **Sentiment Verdict**: Positive
    - **Tone Prediction**: Not Sarcastic
- **Input**: "Wow, great job with another delay! Keep it up."
    - **Sentiment Verdict**: Negative
    - **Tone Prediction**: Sarcastic

This interactive system not only evaluates sentiment but also detects subtle tones in user input, making it an effective tool for customer sentiment analysis.

Conclusion and Limitations

In accordance with our Sentiment Analysis of Airline Tweets, the major work done was with data cleaning and EDA to prepare the data for sentiment classification. We cleaned our dataset by removing irrelevant elements such as stop words, URLs, and user mentions using Twitter data related to major U.S. airlines, and preprocessed each of them to have higher-quality input to analyze.

Some key findings from EDA include the dominance of negative sentiments, correlation between length of tweets and polarity, and recurrence through hashtag and word frequency analysis. We also show that distribution of tweet lengths is approximately symmetric and unimodal with no significant outliers, which will make modeling our data easier in future phases.

These findings will drive feature engineering and modeling at the next phase of work, ensuring that any sentiment prediction made is informed by data. This phase, in general, laid a very good foundation because it gave much insight into the dataset which is a fact that will be crucial in building a strong predictive model.

References

1. Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, Sweta Tiwari,"Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.8, No.4, pp.54-62, 2016. DOI: 10.5815/ijieeb.2016.04.07
2. Yohanssen Pratama et al 2019 J. Phys.: Conf. Ser. 1175 012102
3. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8567243
4. M.Govindarajan , December 2013, Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm
5. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6623713
6. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7011523