Predicting Negative Sentiment in Airline Tweets Using Machine Learning

Patrick Miller & Ganesh Ramcharan

**Introduction:**

Air travel is a popular mode of transportation for long-distance travel. On a daily basis, an average of 1.73 million people use air travel in the United States, but it can be plagued with various issues, such as lost luggage, flight delays, and other problems[1]. This project aims to analyze customer sentiment toward airlines using data mining techniques. Our motivation is to assess how accurately we can predict the most common customer complaints related to airlines so that it can be better for passengers to have a better flying experience. Our goal is to identify the most common negative sentiment towards a specific airline using machine learning algorithms.

We will utilize the Twitter US Airline Sentiment dataset to build supervised machine learning models based on the given features. We will evaluate our models using appropriate metrics, such as precision, recall, and F1 score, to achieve the highest accuracy[3]. We will start by cleaning the data and filling in missing values using the median method. Next, we will use the Naive Bayes algorithm to analyze the data and identify the most common customer complaints and the flights they occurred on[4].

**Description:**

Our dataset for analysis is the Twitter US Airline Sentiment dataset, containing over 14,000 tweets from February 2015 that mention major US airlines including American, Delta, Southwest, United, US Airways, and Virgin America. The dataset is stored in an SQLite database and includes fifteen columns with information about each tweet. Specifically, the text columns include the actual tweet text, while the "airline\_sentiment" column categorizes each tweet as positive, negative, or neutral. The "airline\_sentiment\_confidence" column provides a measure of confidence in the assigned sentiment label as determined by annotators.

To prepare the data for analysis, we will first perform a thorough cleaning process. This will involve removing irrelevant information such as "tweet\_id", "negativereason\_gold", and "retweet\_count". Additionally, we will address any missing data points using the median method to ensure our dataset is complete. Finally, we will apply tokenization to the cleaned text data, breaking it down into individual words and phrases, and converting it into numerical data to enable machine learning algorithms to effectively analyze the text data[5].

**Hypothesis:**

Air travel is a popular mode of transportation for long-distance travel, but it can come with various issues, such as lost luggage, flight delays, and other problems. This project aims to analyze customer sentiment toward airlines using data mining techniques. Our motivation is to assess how accurately we can predict the most common customer complaints related to airlines, with the goal of potentially improving the flying experience for passengers. Specifically, we aim to identify the most common negative sentiments expressed toward a specific airline using machine learning algorithms.

To prove or disprove our hypothesis, we will utilize the Twitter US Airline Sentiment dataset, which consists of over 14,000 tweets from February 2015 mentioning major US airlines including American, Delta, Southwest, United, US Airways, and Virgin America. We will first perform a thorough cleaning process to remove irrelevant information such as "tweet\_id", "negativereason\_gold", and "retweet\_count". Additionally, we will address any missing data points using the median method to ensure that our dataset is complete.

Next, we will apply tokenization to the cleaned text data, breaking it down into individual words and phrases, and converting it into numerical data that can be used as input for machine learning algorithms. We will then use the Naive Bayes algorithm to analyze the data and identify the most common customer complaints and the flights they occurred on. We will evaluate the performance of our models using appropriate metrics, such as precision, recall, and F1 score, to achieve the highest accuracy. Overall, we believe that our project will help airlines address the most common customer complaints and improve the flying experience for passengers.

**Timeline:**

|  |  |  |
| --- | --- | --- |
| **Phase #** | **Deliverables** | **Date** |
| 2 | Visualization of data | 4/19/23 |
| 2 | Prep data for modeling | 4/22/23 |
| 2 | Possible final draft | 4/25/23 |
| 3 | Try to run models and test data | 5/1/23 |
| 3 | Evaluate the data and write out results and conclusion | 5/6/23 |

**Distribution of Work:**

|  |  |  |
| --- | --- | --- |
| **Phase #** | **Deliverables** | **Owner** |
| 1 | Introduction | Pat |
| 1 | Data Description | Pat |
| 1 | Classification & Naïve Bayes | Ganesh |
| 2 | Data Cleansing | Ganesh |
| 2 | Data Visualization | Group |
| 2 | Data Preprocessing | Ganesh |
| 3 | Classification & Naïve Bayes | Pat |
| 3 | Data Modeling | Group |
| 3 | Group Discussion | Group |

**Conclusion:**

Based on our analysis of customer tweets toward airlines using data mining techniques, we will find the most common negative sentiments expressed toward specific airlines. Our goal is to potentially improve the flying experience for passengers by predicting the most common customer complaints related to airlines. By identifying the most common negative sentiments, airlines can take steps to address these issues and ultimately enhance customer satisfaction.

**References**

[1] <https://www.gaytravel.com/gay-blog/airline-and-flight-statistics#:~:text=In%20the%20United%20States%2C%20the,time%2C%20depending%20on%20the%20source>

[2] <https://developer.ibm.com/tutorials/learn-classification-algorithms-using-python-and-scikit-learn/>

[3]<https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd>

[4] <https://developer.ibm.com/tutorials/learn-classification-algorithms-using-python-and-scikit-learn/>

[5] [Tokenization in Python | Methods to Perform Tokenization in Python (educba.com)](https://www.educba.com/tokenization-in-python/)