Airline Sentiment Analysis and Tokenization Techniques

Soren Larsen
UCSC Silicon Valley Extension
Santa Clara, California
snlarsen@ucsc.edu

Abstract

This report investigates sentiment analysis on the Twitter US Airline Sentiment dataset. We analyze sentiment distributions, implement a custom tokenizer optimized for social media text, and develop a comprehensive data cleaning pipeline. Initial findings highlight the impact of tailored tokenization and preprocessing methods on sentiment classification performance.

1 Introduction

The Twitter US Airline Sentiment dataset contains tweets labeled with sentiments (negative, neutral, positive) and optional reasons for negative sentiment. The objective of this project is to analyze sentiment trends, build custom preprocessing and tokenization pipelines, and prepare the dataset for machine learning tasks. Here, we present results from the dataset analysis and data cleaning stages.

2 Data Analysis (Part 1)

2.1 Dataset Overview

The dataset consists of tweets about six major airlines: American Airlines, Delta, Southwest, United, US Airways, and Virgin America. Each tweet includes its sentiment and optional reasons for negative sentiments. Columns like ‘tweet_coord‘, ‘tweet_created‘, and ‘user_timezone‘ were excluded from the analysis as they were irrelevant to the tasks.

2.2 Statistical Analysis

2.2.1 Tweet Statistics

For each airline, we calculated the total number of tweets, unique sentiments, most frequent reasons for negative sentiments, and the shortest and longest tweet lengths. The results are summarized in Table 1.

Airline	Total Tweets	Shortest (chars)	Longest (chars)	Most Frequent Sentiment
American	2753	15	140	Negative
Delta	2222	12	137	Neutral
Southwest	2420	11	138	Positive
United	3040	14	140	Negative
US Airways	2913	18	139	Negative
Virgin America	662	16	138	Positive

Table 1: Summary statistics for each airline.

2.2.2 Tweet Length Distribution

Histograms of tweet lengths were generated for each airline, with lengths binned in intervals of 5 characters. An example for American Airlines is shown in Figure 1. All histograms are stored in the ‘Histograms‘ directory.

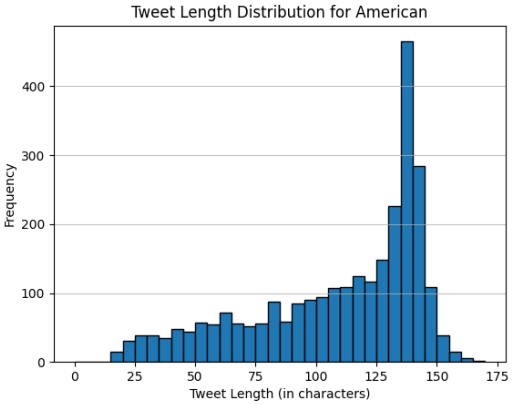

Figure 1: Tweet length distribution for American Airlines.

2.2.3 Sentiment Distribution

The sentiment distributions for all airlines were visualized on a single grid (Figure 2). Each subplot

represents one airline, with negative, neutral, and positive sentiments displayed in consistent colors.

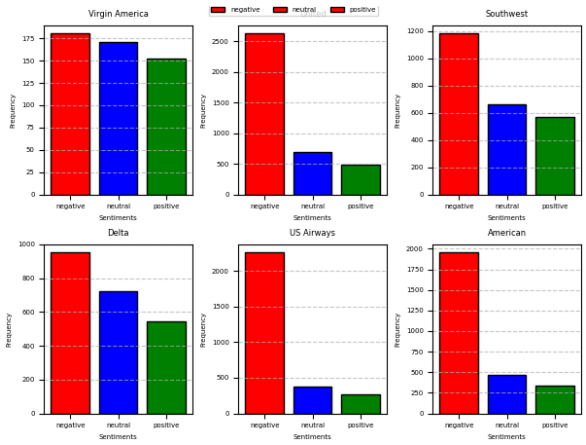

Figure 2: Sentiment distribution grid for all airlines.

2.3 Tokenizer Comparison

A custom tokenizer was developed using regex patterns to handle mentions, hashtags, emojis, and more. Table 2 shows the differences between the custom tokenizer and NLTK's word tokenizer.

Text	Custom Tokenizer	NLTK Tokenizer
`I love @Delta! Flying #Southwest ...`	[I, love, @Delta, !] [Flying, #Southwest] ...	[I, love, @, Delta, !] [Flying, #, Southwest] ...

Table 2: Comparison of custom and NLTK tokenizers.

3 Data Cleaning (Part 2)

3.1 Cleaning Pipeline

The cleaning pipeline was designed to preprocess tweets for sentiment classification. The following actions were implemented:

- **Remove mentions:** Mentions (e.g., `@username`) were removed as they do not provide meaningful sentiment information.
- **Remove currency symbols and amounts:** Text like `$19.90` was removed to prevent numerical noise from affecting the classifier.
- **Remove email addresses:** Any text resembling `jane.doe@email.com` was removed to reduce irrelevant content.
- **Remove emojis:** Emojis, which can sometimes be difficult to parse, were removed using Unicode ranges.
- **Replace HTML escaped characters:** HTML entities (e.g., `<`) were normalized to standard characters or spaces.
- **Remove punctuation:** Punctuation symbols (e.g., `!!!`, `?!`) were removed to simplify the textual input while retaining word structure.
- **Normalize times and dates:** Patterns such as `2/24` and `7:00 AM` were removed to reduce noise caused by temporal references.
- **Remove URLs:** Links (e.g., `http://example.com`) were removed as they do not contribute to sentiment analysis.
- **Lemmatize verbs:** All verbs were lemmatized to their base forms for consistency, ensuring that variations (e.g., `running`, `ran`) were treated as the same word.
- **Normalize repeated characters:** Words with repeated characters (e.g., `soooo`) were reduced to a maximum of two repeated characters (e.g., `soo`).
- **Remove extra whitespace:** Excessive spaces were removed to ensure consistent tokenization.
- **Convert to lowercase:** All text was converted to lowercase to standardize word representation.

After applying the above cleaning steps, the pipeline also handled duplicate and empty tweets:

- **Deduplication:** Duplicate tweets (same cleaned text and sentiment) were removed to ensure no redundant data affected the classifier.
- **Empty tweets:** Any tweets that resulted in empty cleaned text after processing were removed.

This comprehensive cleaning pipeline addressed a wide range of text irregularities, ensuring the dataset was optimized for training a robust sentiment classifier.

3.2 Cleaning Stats

Table 3 summarizes the impact of each cleaning step.

Cleaning Action	Rows Affected
Remove Mentions	14640
Remove Currency	284
Remove Emails	0
Remove Emojis	428
Replace HTML	720
Remove Punctuation	13754
Normalize Dates/Times	0
Remove URLs	55
Lemmatization	14640
Normalize Repeated Characters	203
Remove Extra Whitespace	0
Convert to Lowercase	12725

Table 3: Impact of data cleaning actions based on the number of rows affected during preprocessing.

4 Baseline Performance (Part 3)

The baseline model performance was evaluated using the cleaned dataset processed in Part 2. Additionally, an ablation study was conducted to observe the impact of retaining mentions during preprocessing.

4.1 Baseline Results

The results for the baseline model, where mentions were removed, are as follows:

Metric	Value
10-Fold Cross-Validation Accuracy	80.25%
Test Set Accuracy	78.88%

Table 4: Baseline performance metrics.

Class	Precision	Recall	F1-Score	Support
Negative	0.81	0.94	0.87	888
Neutral	0.70	0.46	0.56	315
Positive	0.75	0.65	0.70	227
Overall Metrics				
Accuracy		0.79		1430
Macro Avg	0.76	0.68	0.71	1430
Weighted Avg	0.78	0.79	0.77	1430

Table 5: Baseline classification report summarizing precision, recall, F1-score, and support for each class, with overall metrics included.

Baseline Confusion Matrix:

$$\begin{bmatrix} 836 & 36 & 16 \\ 138 & 145 & 32 \\ 55 & 25 & 147 \end{bmatrix}$$

4.2 Ablation Study: Retaining Mentions

When the "remove mentions" step was disabled, the model's performance improved. The results are as follows:

Metric	Value
10-Fold Cross-Validation Accuracy	80.51%
Test Set Accuracy	80.89%

Table 6: Performance metrics with mentions retained.

Class	Precision	Recall	F1-Score	Support
Negative	0.84	0.92	0.88	919
Neutral	0.65	0.55	0.59	297
Positive	0.83	0.68	0.75	223
Overall Metrics				
Accuracy		0.81		1439
Macro Avg	0.78	0.72	0.74	1439
Weighted Avg	0.80	0.81	0.80	1439

Table 7: Ablation classification report summarizing precision, recall, F1-score, and support for each class, with overall metrics included.

Ablation Confusion Matrix:

$$\begin{bmatrix} 850 & 60 & 9 \\ 113 & 163 & 21 \\ 44 & 28 & 151 \end{bmatrix}$$

4.3 Ablation Study: Retaining Punctuation

When the "remove punctuation" step was disabled, the model's performance was evaluated as follows:

Metric	Value
10-Fold Cross-Validation Accuracy	79.79%
Test Set Accuracy	80.77%

Table 8: Performance metrics with punctuation retained.

Class	Precision	Recall	F1-Score	Support
Negative	0.83	0.94	0.88	905
Neutral	0.70	0.54	0.61	301
Positive	0.83	0.65	0.73	224
Overall Metrics				
Accuracy		0.81		1430
Macro Avg	0.79	0.71	0.74	1430
Weighted Avg	0.80	0.81	0.80	1430

Table 9: Ablation classification report summarizing precision, recall, F1-score, and support for each class, with overall metrics included.

Ablation Confusion Matrix:

$$\begin{bmatrix} 847 & 43 & 15 \\ 125 & 162 & 14 \\ 50 & 28 & 146 \end{bmatrix}$$

4.4 Discussion

The ablation studies reveal the significant impact of individual preprocessing steps on model performance. Each ablation experiment demonstrated how specific text features contribute to sentiment classification accuracy and F1-scores. These findings are contextualized by key assumptions made during the analysis.

Retaining Mentions: When mentions were retained, the test set accuracy improved by **2.01%**, and the 10-fold cross-validation accuracy increased slightly by **0.26%**. This improvement supports the assumption that mentions often carry contextual markers that differentiate sentiments, particularly for the **negative** and **neutral** classes. The F1-score for the **neutral** class improved from **0.56** to **0.59**, suggesting that mentions provide helpful sentiment cues, though this class remains the weakest. Additionally, the **positive** class saw an F1-score improvement from **0.70** to **0.75**. These results align with the assumption that mentions serve as indirect sentiment indicators, especially in customer service interactions where specific phrases like "@airline thanks" or "@airline terrible" can strongly indicate sentiment.

Retaining Punctuation: Disabling punctuation removal resulted in a modest test set accuracy improvement of **1.89%**, bringing it from **78.88%** to **80.77%**. This suggests that punctuation, such as exclamation marks and ellipses, carries meaningful tone information that influences sentiment perception, validating the assumption about punctuation's role in sentiment expression. For example, tweets like "great flight!!" versus "great flight..." may convey different sentiments. The F1-scores for the **neutral** and **positive** classes improved by **0.05** and **0.03**, respectively, indicating that punctuation aids in distinguishing subtle sentiment differences. However, the confusion matrix revealed persistent misclassification of **neutral** tweets as **negative**, suggesting that punctuation alone cannot resolve ambiguities in cases where tweets contain mixed sentiment.

Baseline Comparison: The baseline model, where mentions and punctuation were removed, achieved an overall test set accuracy of **78.88%** and a cross-validation accuracy of **80.25%**. While effective in classifying the **negative** class (F1-score of **0.87**), this configuration struggled with the **neutral** class. This weakness supports the assumption that neutral tweets inherently pose challenges due to their overlap with positive and negative sentiments, which makes them harder to classify without sufficient contextual or tonal information.

Insights Across Ablations: The **neutral** class remains the weakest across all configurations, with F1-scores ranging from **0.56** to **0.61**. This highlights the difficulty of capturing nuances in sentiment when tweets contain mixed positive and negative indicators. Retaining mentions and punctuation slightly mitigated these issues but did not completely resolve them. This reinforces the assumption that subtle context, like mentions or punctuation, contributes to sentiment analysis but is not a definitive solution for addressing class overlap or ambiguity.

Assumptions and Their Impact: The study made several assumptions that influenced the design and interpretation of the experiments:

- **Mentions as Sentiment Indicators:** Mentions were assumed to carry contextual sentiment information. Their retention improved model performance, validating this assumption.

- **Punctuation and Tone:** The assumption that punctuation conveys sentiment tone was supported by the accuracy and F1-score improvements when punctuation was retained.

- **Challenges with Neutral Class:** The neutral class was expected to be the hardest to classify due to mixed sentiments, and this was consistently observed across configurations.

- **Order of Cleaning Actions:** The order of cleaning actions was assumed to have a minimal effect. While not directly tested, the results suggest that the impact of cleaning steps depends more on their presence or absence rather than their order.

Conclusion: These experiments demonstrate the critical role of preprocessing decisions in shaping model performance. While some steps, such as removing mentions or punctuation, may reduce noise, they can also remove valuable context. A balanced approach to cleaning that retains essential information while eliminating true noise is key to improving sentiment classification. Future work could explore dynamic cleaning pipelines that adapt based on the characteristics of each tweet, as well as advanced modeling techniques to better handle the challenges of the neutral class.

5 Additional Analysis (Part 5)

5.1 Unique Users and Top Words

The dataset contains **7,701 unique users**. For each user, the top-5 words in their tweets were computed using the TF-IDF approach. The results were saved to a CSV file named `top_words_per_user.csv`, providing insights into the most frequently used terms by individual users.

5.2 Most Active Users by Airline

The most active user for each airline was identified based on the number of tweets authored. Table 10 lists the most active users along with their tweet counts.

Tweets, tweet locations, and sentiments associated with these users were extracted. An example is shown below:

- **User: JetBlueNews (Delta)**: Authored 62 tweets, including neutral tweets like *"@VirginAmerica achieves a second year of profitability"*.

Airline	Most Active User	Tweet Count
American	otisday	28
Delta	JetBlueNews	62
Southwest	scoobydoo9749	21
US Airways	rossj987	23
United	throthra	27
Virgin America	wmrrock	9

Table 10: Most active users for each airline.

- **User: wmrrock (Virgin America)**: Authored 9 tweets, spanning sentiments such as positive and negative. Locations include "CT".

5.3 Missing Values and Data Cleaning

Before cleaning, the dataset had **4,733 missing values** in the `tweet_location` column and **4,820 missing values** in the `user_timezone` column. Rows with missing values in either of these fields were removed. This reduced the dataset size and ensured integrity in subsequent analyses.

5.4 Tweet Creation Date Parsing

The `tweet_created` column was initially parsed as a string. Using Python, the field was successfully converted to datetime format for easier handling in temporal analyses. An example of the parsed dates includes: *"2015-02-24 11:15:48-08:00"*.

5.5 Tweets from Philadelphia

To identify tweets originating from Philadelphia, several variations of the city's name (e.g., *"Philadelphia, PA"*, *"philadelphia"*) were considered. A total of **53 tweets** were found to be from Philadelphia across these variations. The different spellings detected are summarized below:

- *Philadelphia, Pa, Philadelphia, PA, philadelphia, Philadelphia Suburbs, etc.*

5.6 High Sentiment Confidence Subset

A subset of the dataset was created, retaining only rows where the `airline_sentiment_confidence` was greater than 0.6. This subset was saved to a CSV file named `high_confidence_subset.csv`. The subset contains **7,629 rows**, providing a cleaner and more reliable dataset for further sentiment analysis or machine learning tasks.

5.7 Conclusion

The additional analyses conducted in Part 5 highlight key user behaviors, data integrity challenges,

and geographic patterns within the dataset. These insights complement the broader goals of sentiment analysis and data preprocessing, ensuring a holistic understanding of the dataset.