NLP 220 Assignment 3

sttau

December 2023

1 Part 1

1.1 Part A

The dataset comprises a total of 13640 tweets targeting six major airlines. The breakdown of the number of tweets per airline is as follows:

- United: 3822
- US Airways: 2913
- American: 2759
- Southwest: 2420
- Delta: 2222
- Virgin America: 504

The length of tweets ranges from 12 to 186 characters.

1.1.1 Virgin America

Column	Number of Unique Values	Most Frequent Value
airline_sentiment	3	Negative (181 occurrences)
negativereason	10	Customer Service Issue (60 occurrences)

Table 1: Sentiment Analysis for Virgin America

1.1.2 United

Column	Number of Unique Values	Most Frequent Value
airline_sentiment	3	Negative (2633 occurrences)
negativereason	10	Customer Service Issue (681 occurrences)

Table 2: Sentiment Analysis for United

1.1.3 Southwest

Column	Number of Unique Values	Most Frequent Value
airline_sentiment	3	Negative (1186 occurrences)
negativereason	10	Customer Service Issue (391 occurrences)

Table 3: Sentiment Analysis for Southwest

1.1.4 Delta

Column	Number of Unique Values	Most Frequent Value
airline_sentiment	3	Negative (955 occurrences)
negativereason	10	Late Flight (269 occurrences)

Table 4: Sentiment Analysis for Delta

1.1.5 US Airways

Column	Number of Unique Values	Most Frequent Value
airline_sentiment	3	Negative (2263 occurrences)
negativereason	10	Customer Service Issue (811 occurrences)

Table 5: Sentiment Analysis for US Airways

1.1.6 American

Column	Number of Unique Values	Most Frequent Value
airline_sentiment	3	Negative (1960 occurrences)
negativereason	10	Customer Service Issue (768 occurrences)

Table 6: Sentiment Analysis for American

1.2 PartB

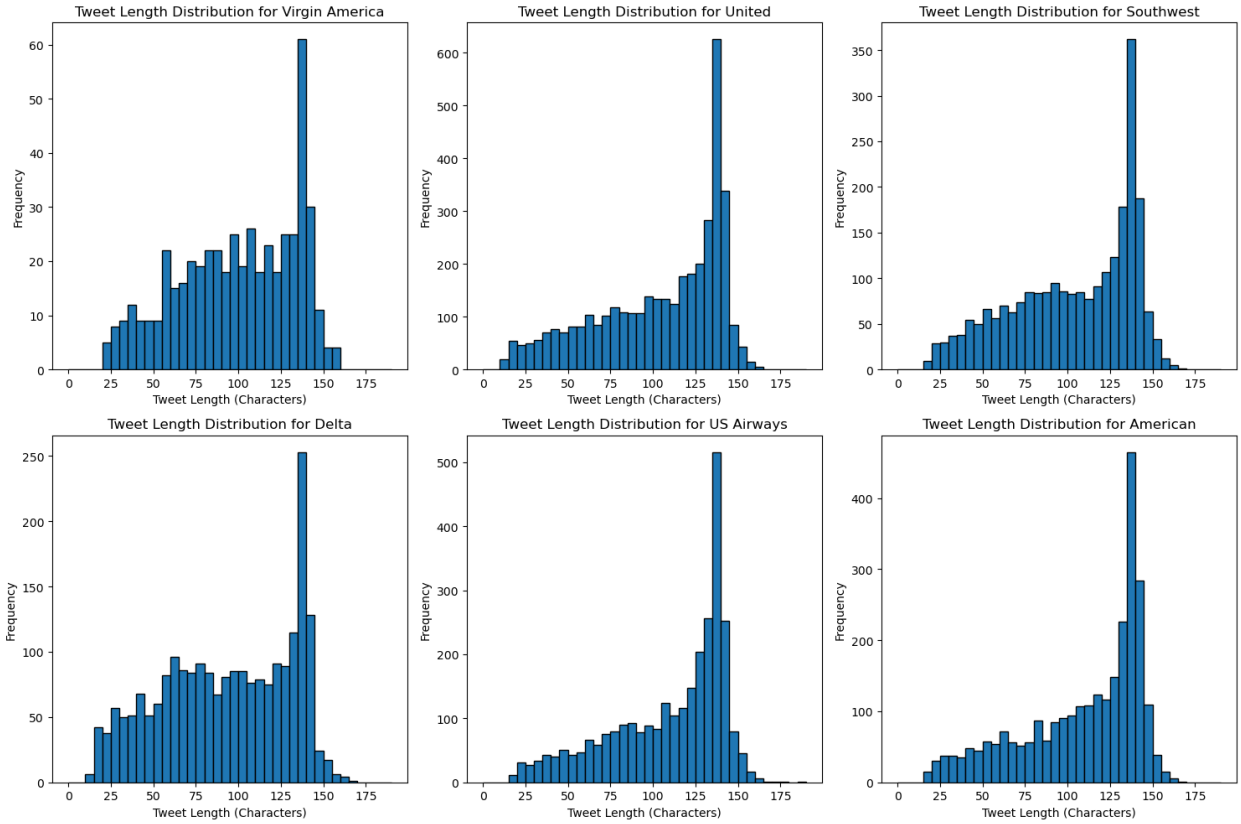

Figure 1: Tweet Length across airlines.

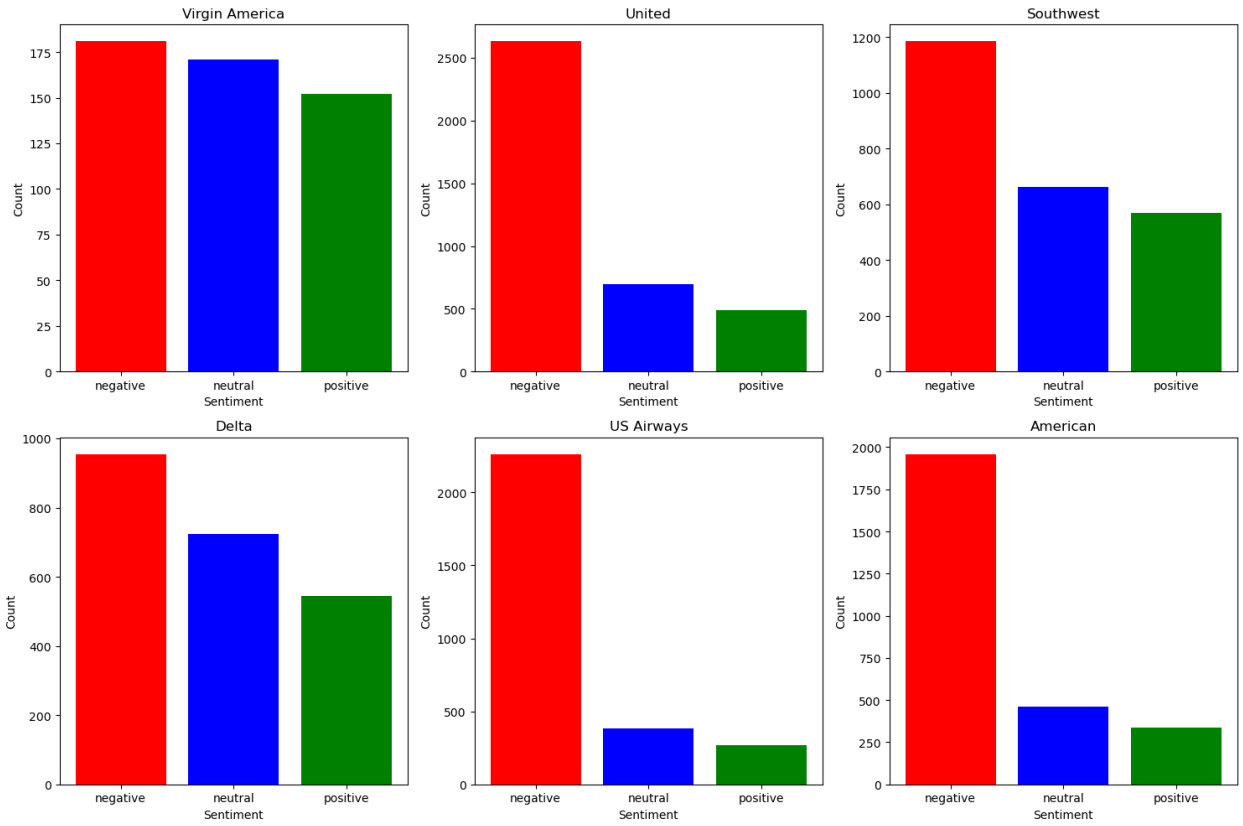

Figure 2: Sentiment across airlines.

1.3 Part C

The following ordered list describes each step in the tweet processing pipeline:

1. **Process HTML Entities:** Converts HTML entities in the text into their corresponding characters. For example, `<` becomes <, and `&` becomes &.

2. **Remove Non-Standard Characters:** Filters out characters that are not standard English characters, numbers, or basic punctuation. It helps in cleaning up the text by removing unwanted symbols or characters.

3. **Process Emoticons:** Converts emoticons into a textual representation or removes them, depending on the implementation. It's essential for capturing the sentiment or tone conveyed by emoticons.

4. **Tokenize URLs, Hashtags, and Mentions:** Identifies and handles URLs, hashtags, and mentions (`@username`) in the text, potentially replacing them with placeholders or handling them specially.

5. **Lowercase Tokens (Excluding Processed Emoticons):** Converts the text to lowercase to ensure consistency, except for parts of the text that have been processed as emoticons. This step is standard in text processing to reduce variability caused by case differences.

1.4 Part D

NLTK processes punctuation as its own tokens while mine doesn't. I do process HTML entities and emoticons to their respective expressions. I forgot to split on punctuation and could of handle basic cases with apostrophes and periods.

Example	NLTK Tokens	Custom Tokens
1	['What', 'said', '.']	['what', 'said.']
2	['plus', 'you', "ve', 'added', 'commercials', 'to', 'the', 'experience', '...', 'tacky', '.']	['plus', 'you've', 'added', 'commercials', 'to', 'the', 'experience...', 'tacky.']
3	['I', 'did', 'n't', 'today', '...', 'Must', 'mean', 'I', 'need', 'to', 'take', 'another', 'trip', '!']	['i', 'didn't', 'today...', 'must', 'mean', 'i', 'need', 'to', 'take', 'another', 'trip!']
4	['it', "s', 'really', 'aggressive', 'to', 'blast', 'obnoxious', '"', 'entertainment', '"', 'in', 'your', 'guests', "', 'faces', ' '	["it's", 'really', 'aggressive', 'to', 'blast', 'obnoxious', '"entertainment"', 'in', 'your', "guests"', 'faces', ', 'amp', ';', 'they', 'have', 'little', 'recourse']
'	', 'they', 'have', 'little', 'recourse']	
5	['and', 'it', "s', 'a', 'really', 'big', 'bad', 'thing', 'about', 'it']	['and', "it's", 'a', 'really', 'big', 'bad', 'thing', 'about', 'it']

Table 7: Comparison of NLTK and Custom Tokenization Methods

2 Part 2

Included the prior 5 preprocessing from the previous ppiline

1. **Clean Currency**: Standardize or remove currency representations in the text.
2. **Clean Emails**: Remove or anonymize email addresses to reduce noise and protect privacy.
3. **Clean Emojis**: Convert emojis into textual representations for interpretability by text analysis models.
4. **Clean Punctuation**: Normalize punctuation marks to maintain sentence structure and reduce noise.
5. **Clean Dates and Times**: Standardize or remove date and time references, unless specifically required for the analysis.
6. **Clean URLs**: Remove or replace URLs, as they typically do not contribute to the overall sentiment or content analysis.
7. **Clean Numbers**: Process numerical data by removing or normalizing numbers.
8. **Normalize Whitespace**: Reduce multiple whitespace characters to a single space for consistent spacing.
9. **Lemmatize Verbs**: Convert verbs to their base form to simplify the text for easier processing by models.

3 Part 3

The mean cross-validation accuracy of the model is 0.8015.

The model achieved a classification accuracy of 0.7823 on the test set.

The classification report for the model is as follows:

```
              precision    recall  f1-score   support

    negative       0.80      0.93      0.86       886
     neutral       0.69      0.45      0.55       324
    positive       0.78      0.67      0.72       223

    accuracy                           0.78      1433
   macro avg       0.76      0.68      0.71      1433
weighted avg       0.77      0.78      0.77      1433
```

The confusion matrix for the model's predictions is:

	Negative	Neutral	Positive
Negative	826	43	17
Neutral	153	146	25
Positive	52	22	149

Preprocessing Step	Performance (Accuracy)
Clean Currency	0.7865
Clean Emails	0.7864
Clean Emojis	0.7892
Clean Punctuation	0.7857
Clean Dates and Times	0.7866
Clean URLs	0.7863
Clean Numbers	0.7880
Clean Hashtags	0.7846
Process Emoticons	0.7880
Remove Non-Standard Characters	0.7862
Normalize Whitespace	0.7864
Lemmatize Verbs	0.7850

Table 8: Model Performance with Individual Preprocessing Steps

For ablation, the individual components did not contribute that much to the over-test set as all of them combined. This suggests that preprocessing might not be needed for the SVM model. It is possible certain combinations could contribute a significantly higher score for the final preprocessing step.