Data Appendix

DS 4002

Group 11: Ctrl + Alt + Elite

Rhea Hemrajani (Group Leader), Kaitlin Blakeslee, Carson Smith

1. **Top Words with Highest TF-IDF Scores**
   Unit of observation: individual tweets

```
Top 10 most predictive words:
    cancelled    0.103450
    delayed      0.103328
    service      0.067655
    customer     0.017107
    flightled    0.004695
    amp          0.000288
    usairways   -0.000601
    hours       -0.001736
    flights     -0.003769
    plane       -0.009133
```

- x
    - Represents the independent variables (features) used for training the linear regression model. In this case, it contains the TF-IDF scores of words from df_combined. X is used to predict the negative sentiment scores based on the word frequencies in the tweets.
    - Derived from df_combined by dropping the sentiment_column.
- y
    - Represents the dependent variable (target) that the linear regression model aims to predict. In this case, it contains the negative sentiment scores from df_combined. y is what the model tries to predict based on the values in X.

2. **Top Predictive Words for AIrline Negative Sentiment**

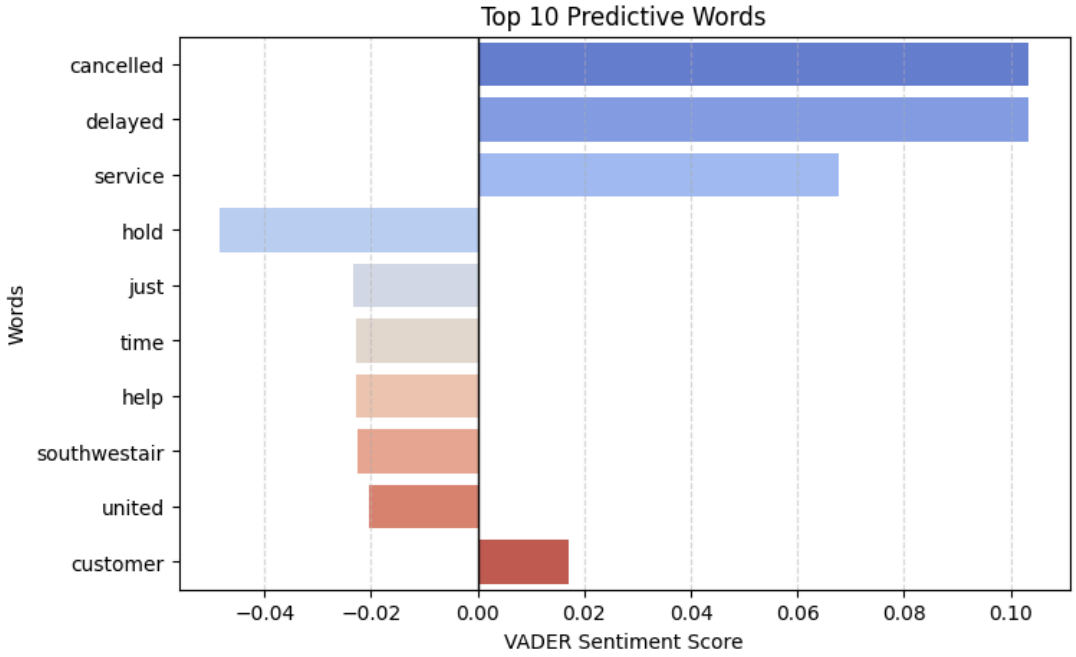
Unit of Observation: individual words in tweets

Feature_importance

- This is the most important variable for the word cloud. It's likely a dictionary or a Pandas Series where higher values for a word in feature_importance mean that word is a stronger predictor of negative sentiment and will appear larger in the word cloud.
    - feature_importance provides the data (words and their importance), wordcloud processes it to create the word cloud image, and plt (matplotlib) handles displaying that image. Each word's size within the cloud is directly proportional to its value in feature_importance, highlighting the most predictive words for negative sentiment.

2. **Top 10 Predictive Words**

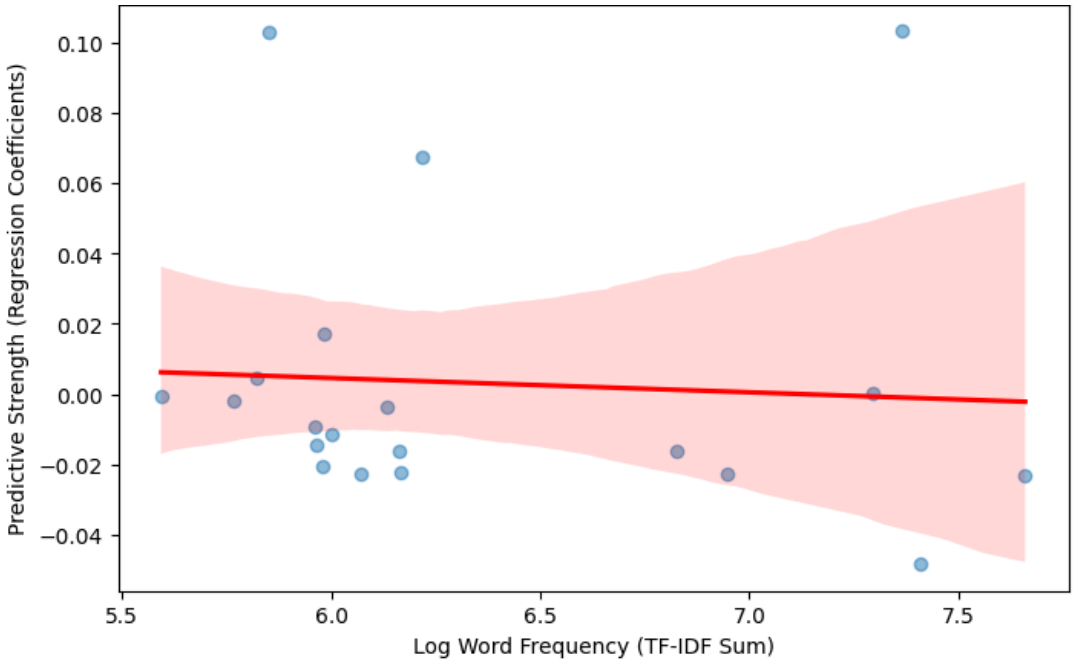
Unit of observation: individual tweets



- Word
  - Represents the actual words that were identified as the top 10 predictors of negative sentiment.
  - Each bar in the chart corresponds to a different word.
  - plotted on the y-axis
- VADER Sentiment Score**:**
  - Represents the strength and direction of the relationship between each word and negative sentiment – essentially a measure of how much each word contributes to predicting negative sentiment.
  - The further a bar extends to the right (positive values), the more strongly that word is associated with negative sentiment. Bars extending to the left (negative values) indicate words that might be associated with less negative or even positive sentiment.
  - Plotted on the x-axis of the bar chart.

3. **Log-Transformed Relationship Between Word Frequency and Predictive Strength**

Unit of Observation: individual words in tweets



The code snippet aims to visualize the relationship between word frequency (how often a word appears in the dataset) and its predictive strength (how strongly it's associated with negative sentiment). It does this by creating a scatter plot with a regression line. The log transformation is applied to word frequency to improve visualization. By looking at the plot, you can get an idea of whether words that appear more frequently are also more likely to be associated with negative sentiment.

Log Word Frequency (TF-IDF Sum)**:**

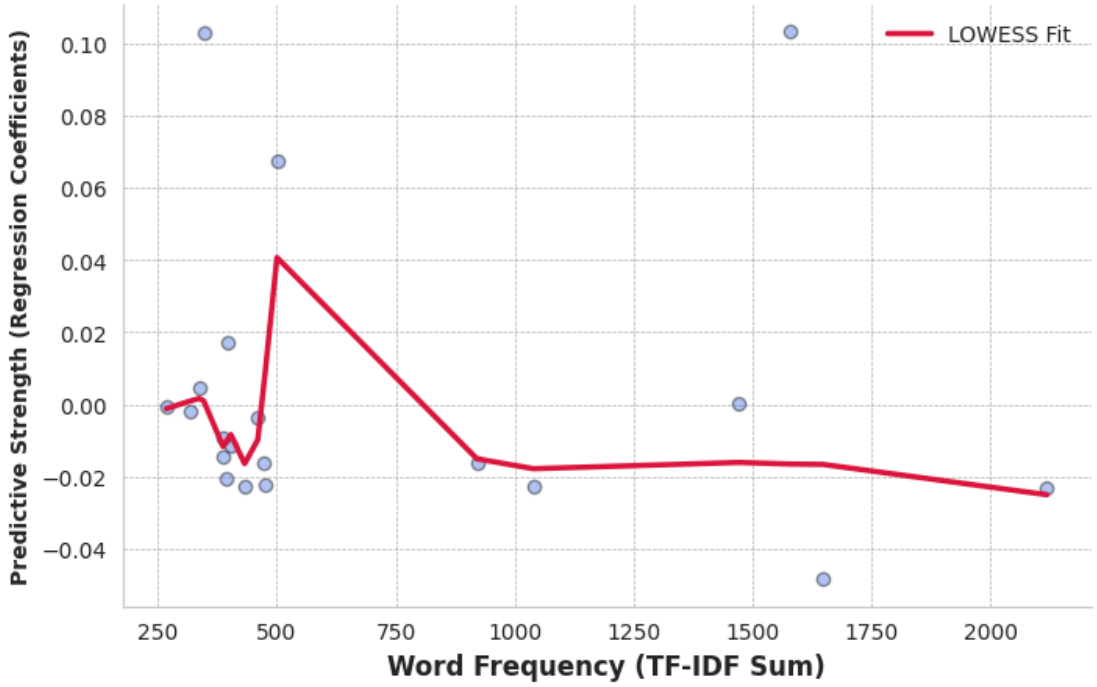
- Shows how often a word appears in the dataset of tweets.
    - tfidf_df.sum() calculates the total frequency of each word across all tweets.
    - np.log1p() is applied to this sum, which means the x-axis values are actually the natural logarithm of (1 + word frequency).
    - Plotted on the X-axis.

Predictive Strength (Regression Coefficients):

- **S**hows how strongly each word predicts negative sentiment in the tweets.
    - Higher values on the y-axis indicate that a word is more strongly associated with negative sentiment.
    - Plotted on the Y-axis

4. **Relationship Between Word Frequency and Predictive Strength**

Unit of Observation: individual words in tweets



The visualization aims to show the relationship between these two variables: Does the frequency of a word affect how strongly it predicts negative sentiment? The LOWESS smoothing technique helps in visualizing this relationship by creating a smooth curve that represents the general trend in the data.

Word Frequency (TF-IDF Sum):

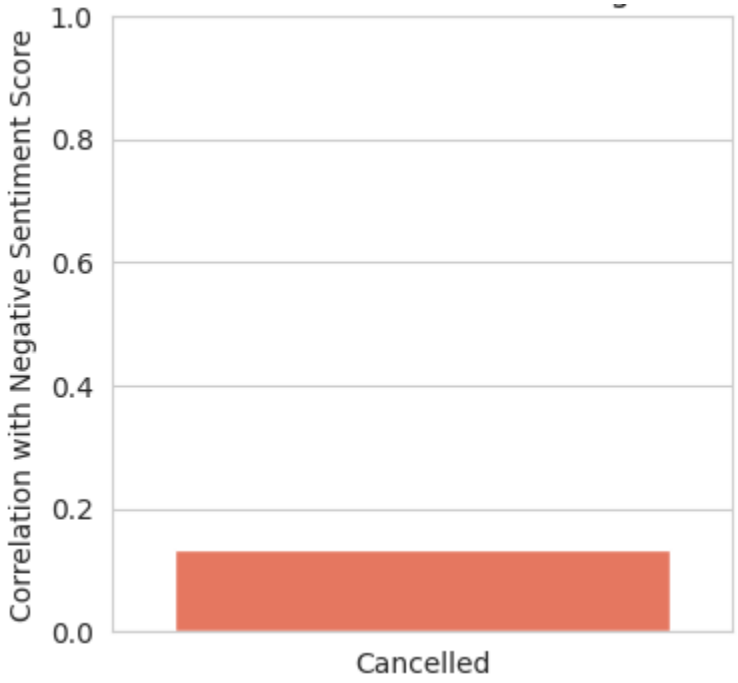
- Represents the word frequency in the dataset.
    - Represents the independent variable, suggesting that word frequency might influence the predictive strength of a word.
    - tfidf_df is a DataFrame where each column represents a word and each row represents a tweet. The cells contain TF-IDF scores, which indicate the importance of a word in a document (tweet) relative to a collection of documents (all tweets).
    - sum() is applied to tfidf_df to calculate the total TF-IDF score for each word across all tweets. This essentially represents how frequently each word appears in the entire dataset.
    - Plotted on x-axis

Predictive Strength (Regression Coefficients)

- Represents the predictive strength or importance of each word in predicting negative sentiment.
    - Higher absolute values of these coefficients indicate stronger predictive power (either positive or negative).
    - It represents the dependent variable, suggesting that the predictive strength of a word might be dependent on its frequency.
    - Plotted on the y-axis.

5. Correlation Between ‘Cancelled’ and Negative Sentiment

Unit of observation: individual tweets



The bar chart visually represents the relationship between the presence of the word "cancelled" in tweets and the negative sentiment expressed in those tweets. The height of the bar is directly determined by the cancelled_corr variable, which quantifies the strength and direction of this relationship. The visualization makes it easy to understand whether the word "cancelled" is a strong indicator of negative sentiment in the analyzed tweets.

“Cancelled”

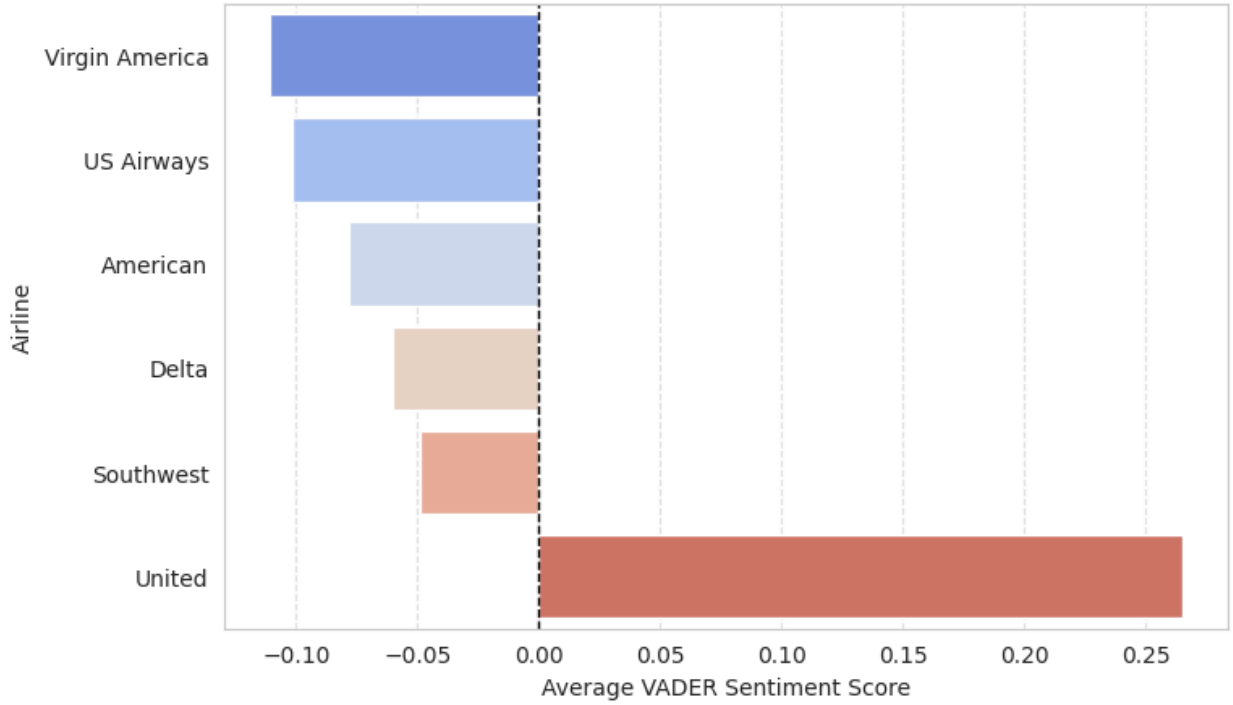
- Represents the category being analyzed. This category represents the presence of the word "cancelled" within the tweets.
    - Plotted on x axis

Correlation with Negative Sentiment Score

- Represents the numerical value associated with the category on the x-axis. The height of the bar corresponds to the correlation value calculated between the presence of the word "cancelled" and the negative sentiment scores of the tweets (0-1)
    - 1 represents a high correlation
    - 0 represents a low correlation
    - Plotted on y axis

6. Average Sentiment Scores Across Airlines

Unit of Observation: Individual tweets



The visualization is a bar chart that shows the average sentiment scores for different airlines. The key variables for the visualization include the following:

Average VADER Sentiment Score

- This is a continuous variable representing the sentiment score calculated by VADER for each tweet. It is essentially the average sentiment score.
    - VADER assigns a score between -1 and 1 to each tweet:
        - Scores closer to -1 indicate negative sentiment.
        - Scores closer to 1 indicate positive sentiment.
        - Scores near 0 indicate neutral sentiment.
    - Represented on x-axis

Airline

- This is a categorical variable representing the different airlines mentioned in the tweets. Each airline is a distinct category and is displayed on the y-axis of the bar chart.
    - Represented on y-axis