Data Appendix

The purpose of this appendix is to provide a detailed description of the datasets used in the Exploratory Data Analysis (EDA) and Sentiment Data Analysis. It details information about the variables of the original and cleaned dataset, including descriptive statistics.

Appendix A. Original Dataset (nfl_sentiments_original.csv)

We used a dataset called 'NFL Twitter Sentiment Analysis' (sourced from Kaggle), which is saved as nfl_sentiments_original.csv in the DATA folder. Each unit of observation represents a tweet about a particular NFL team in the form of text data posted by a Twitter user, with the sentiment of the post towards that team. Notably, each tweet in the dataset is posted by a unique user, meaning no twitter accounts are repeated across the entire dataset. There are a total of 5171 entries collected from November and December of 2024 in this original dataset.

Dataset Variables:

<u>username</u>: string variable
- Indicates the unique username of the Twitter user that published the associated tweet.
  - Values: E.g., Gary

<u>timestamp</u>: datetime variable
- Includes the date and time when the tweet was published in UTC format.
  - Values: E.g., 2024-11-29T23:30:57.000Z

<u>text</u>: string variable
- Text data containing the actual contents of the original tweet.
  - Values: E.g., That's a Cleveland Browns type way to lose a game.

<u>team</u>: string variable
- Indicates the team that the unique user's tweet is about.
  - Values: Miami Dolphins; Denver Broncos; Indianapolis Colts; Cleveland Browns; New York Jets; Jacksonville Jaguars; Kansas City Chiefs; Detroit Lions

<u>sentiment</u>: string variable
- Displays the classified sentiment of the unique user's tweet.
  - Values: positive; neutral; negative

<u>confidence</u>: float variable
- States the confidence (certainty score) of the model used by the authors of the dataset to classify the sentiment of each unique user's tweet.
  - Range of Values: 0-1 (e.g., 0.69503664970398)

<u>roberta_raw_outputs</u>: list variable
- States the raw probability scores of each type of sentiment using the RoBERTa model.
  - Values: E.g., [0.6950366497039795, 0.173924520611763, 0.13103878498077393]

Descriptive Statistics:

- Total Entries: 5171
- Range (dates): 11/23/2024–12/17/2024

Appendix B. Final Dataset (nfl_sentiments_cleaned_final_version.csv)

This data file contains the cleaned dataset that we used for our logistic regression analysis to answer our research question: whether fans' pregame sentiment towards various NFL teams have a statistically significant relationship with the success of those teams in the 2024 season. Each unit of observation represents a tweet about a NFL team in the form of text data posted by a Twitter user, with the sentiment of the post towards that team and whether they won or lost. Tweets that occurred outside the range of 11/29/2024–11/30/2024 were filtered out of this dataset to perform our analysis. An additional variable was added to display whether the team won or lost a specified game.

Cleaned Dataset Variables:

username: string variable

- Indicates the unique username of the Twitter user that published the associated tweet.
    - Values: E.g., Rick Ferguson

timestamp: datetime variable (python date class)

- Includes the date and time when the tweet was published in UTC format.
    - Values: E.g., 2024-11-29 23:15:29+00:00

text: string variable

- Text data containing the actual contents of the original tweet.
    - Values: E.g., That's a Cleveland Browns type way to lose a game.

team: string variable

- Indicates the team that the unique user's tweet is about.
    - Values: Miami Dolphins; Denver Broncos; Indianapolis Colts; Cleveland Browns; New York Jets; Jacksonville Jaguars; Kansas City Chiefs; Detroit Lions

sentiment: string variable

- Displays the classified sentiment of the unique user's tweet.
    - Values: positive; neutral; negative

confidence: float variable

- States the confidence (certainty score) of the model used by the authors of the dataset to classify the sentiment of each unique user's tweet.
    - Range of Values: 0-1 (e.g., 0.69503664970398)

roberta_raw_outputs: list variable

- States the raw probability scores of each type of sentiment using the RoBERTa model.
    - Values: E.g., [0.6950366497039795, 0.173924520611763, 0.13103878498077393]

winner:

- Indicates whether the team associated with the post won or lost the game from week 13 of the NFL season.
    - Values: 0, 1
        - '0' represents a win
        - '1' represents a loss

Descriptive Statistics:

- Total Entries: 2026
- Range (dates): 11/29/2024–11/30/2024

After cleaning our dataset, Table 1 displays the number of entries per team in our dataset. Clearly, the Miami Dolphins, Kansas City Chiefs, and Detroit Lions have significantly more data entries than the other teams, which must be considered when interpreting our results. Eventually, the sentiments will be converted to an average sentiment score per team in a dataframe to perform our analysis.

Team	Frequency
Miami Dolphins	778
Kansas City Chiefs	593
Detroit Lions	518
New York Jets	63
Jacksonville Jaguars	42
Colts	32

Table 1. Number of Entries Per Team

The following figures display the general distribution and trends of our dataset. Figure 1 displays the distribution of sentiment scores across all teams, which generally follows a normal distribution while slightly skewing towards the right (i.e., towards positive sentiment). This trend follows similarly in Figure 2, clearly showing a higher frequency of positive sentiment tweets in our dataset overall. Figure 3 further expands on this by showing the most frequently used words across all three types of sentiments. This Word Cloud gives us a visual of the content of the tweets overall as well as showcases how certain teams have more data than others.

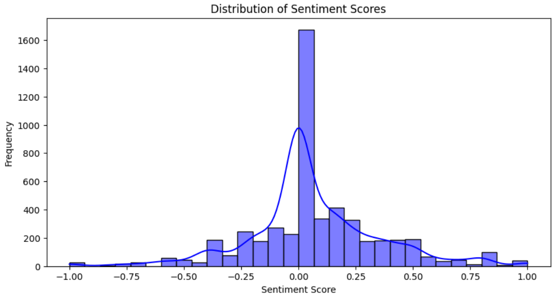

Figure 1. Distribution of Sentiment Scores

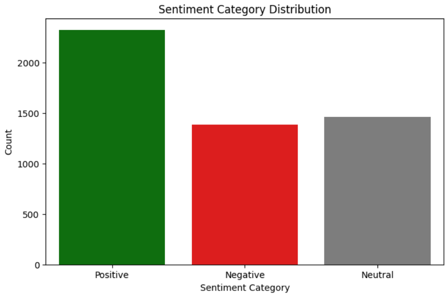

Figure 2. Distribution of Sentiment Categories

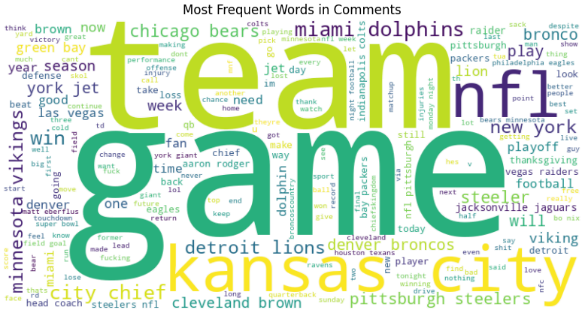

Figure 3. Word Cloud of Most Frequent Words Used in Tweets