CS415 Project 3 Report

Jacob Barkovitch
Binghamton University
Vestal, New York, USA
jbarkov1@binghamton.edu

Guy Ben-Yishai
Binghamton University
Vestal, New York, USA
gbenyis1@binghamton.edu

Alan Bixby
Binghamton University
Vestal, New York, USA
abixby1@binghamton.edu

Jacob Coddington
Binghamton University
Vestal, New York, USA
jcoddin1@binghamton.edu

Ryan Geary
Binghamton University
Vestal, New York, USA
rgeary1@binghamton.edu

Joseph Lieberman
Binghamton University
Vestal, New York, USA
jliebe12@binghamton.edu

Figure 1: The Pie Chart Gambit [1]

Abstract

This report presents a preliminary analysis of the correlation between sentiment data from social media platforms, including Twitter and Reddit, and the outcomes of National Football League (NFL) games. The study uses sentiment analysis and posting frequency data to attempt to predict the expected winner of a game and identify team rivalries and classify reactions to game outcomes. The data used in the analysis covers the period from November 1st to November 27th, 2022 and includes sentiment data hydrated using TextBlob and tweet mapping to teams using player roster data and tweet entitlement ids. The report also includes an OpenAPI-compliant interface using FastAPI and a NextJS-based dashboard to showcase some of the results. However, due to time constraints, a significant portion of the functionality is only accessible through API routes.

1 Introduction

In this report, we expand the preliminary results of Project 2, and craft a preliminary quantitative analysis of the correlation of sentiment data with game out. hen we slap that into a OpenAPI compliant interface using FastAPI, and produce a half-baked NextJS based dashboard to showcase some of the the graphs as dashboards. Unfortunately due to time constraints, a significant portion of the functionality is still locked behind the API routes only.

Significant preliminary analysis was conducted in our Project 2 report, and for the sake of brevity, findings made in the previous report have been omitted from this report.

1.1 Research Question

- Can sentiment analysis and posting frequency of social media data be used to predict the expected winner of a game?

1.2 Related Work

In recent years, researchers have explored the use of sentiment analysis in predicting sports betting outcomes. For example, in 2017, researchers at the University of Texas and Central Connecticut State University[2] demonstrated the ability to profit from NFL sports betting by analyzing sentiment data from tweets. Additionally, a study published in the Journal of Gambling Studies[3] examined the gambling behavior of sports fans and found significant differences in attitudes between sports bettors, non-sports bettors, and non-bettors. These findings align with the expected demographics of users who follow the NFL on social media platforms such as Reddit and Twitter. These studies highlight the potential value of sentiment analysis in understanding and predicting sports betting behavior.

2 Methodology & Dataset

2.1 Data Constraints

The data collected for this report covers the period from noon on November 1st, 2022 to midnight on November 27th, 2022 UTC. During this collection period, our Twitter data suffered data losses from November 8th to November 14th, and November 18th to November 19th; the Odds API data and Reddit data were not affected by these outages.

We will focus on analyzing data from the National Football League (NFL) using Twitter context annotations. While we possess data for all sports, the NFL is particularly well-suited for our analysis because it is currently in-season and is well-defined within the Twitter context annotations. NFL players are well-known and established, allowing for a more precise analysis compared to NCAAF players, who were often missed by the context annotations labelling, likely because they are too young or new to be added to Twitter's dataset.

Lastly, during Project 2 we discovered that the *Los Angeles Rams* subreddit was omitted from our data collection subreddit list. As such, any queries made to the FastAPI endpoints, or analysis in this report will be excluding data pertaining to the *Los Angeles Rams*.

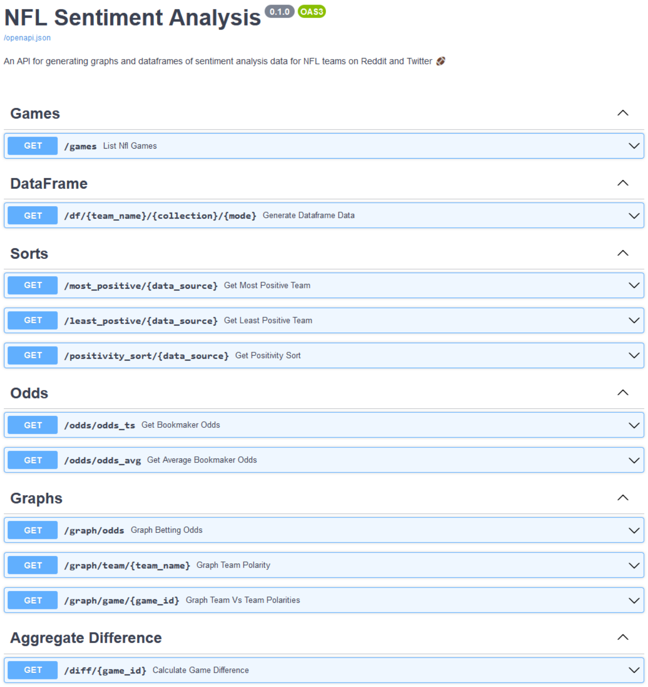
2.2 Data Preparation

Our data was hydrated with sentiment data using TextBlob and tweets were mapped to teams using player roster data and tweet entitlement ids as outlined the Project 2; see teamToEntitlementIds.json in **project3-implementation** for entity id to team mappings. We also fetched the end times/duration of NFL games using pro-football-reference.com as implemented in the nfl-endtime branch. This was intended to be used for team rivalry analysis; but due to time restraints is almost exclusively used in enhancing the plotting visuals. Reddit data is still classified using only the associated team subreddits; see teamToSubreddits.json in **project3-implementation** for subreddit to team mappings.

3 FastAPI

Our primary deliverable is a REST API using FastAPI; our endpoints are categorized into Games, DataFrames, Sorts, Odds, Graphs, and Aggregate Difference. Utilizing the **project3-implementation** README.md likely will be the most accurate source of response schemas and parameter descriptions, due to the lackluster native support for inferring mypy types and a lack of time to create a proper yaml config or Pydantic schema.

Alternatively, **/df/{team_name}/{collection}/{mode}** can be used to fetch the raw data in a JSON format that is used to generate the metrics used in the subsequent endpoints. All of the following is accessible at https://api.ds.bxb.gg/docs in OpenAPI format, as pictured in Figure 2.



NFL Sentiment Analysis 0.1.0 OAS3

/openapi.json

An API for generating graphs and dataframes of sentiment analysis data for NFL teams on Reddit and Twitter 🏈

Games

GET /games List Nfl Games

DataFrame

GET /df/{team_name}/{collection}/{mode} Generate Dataframe Data

Sorts

GET /most_positive/{data_source} Get Most Positive Team

GET /least_postive/{data_source} Get Least Positive Team

GET /positivity_sort/{data_source} Get Positivity Sort

Odds

GET /odds/odds_ts Get Bookmaker Odds

GET /odds/odds_avg Get Average Bookmaker Odds

Graphs

GET /graph/odds Graph Betting Odds

GET /graph/team/{team_name} Graph Team Polarity

GET /graph/game/{game_id} Graph Team Vs Team Polarities

Aggregate Difference

GET /diff/{game_id} Calculate Game Difference

Figure 2: OpenAPI Documentation Page

3.1 Games

3.1.1 /games

For ease of use, **/games** provides a JSON dump of the *nfl-games* collection from MongoDB to quickly see *game ids* and outcomes; the *game id* will be relied on for several subsequent endpoints. This endpoint is also a useful summary for insights on the games within the context of our research.

3.2 Graphs

3.2.1 /graph/team/{team_name}

Provided a *game id* and a target data source (Reddit or Twitter), this endpoint generates a graph of the community sentiment as a rolling average of defined by the query parameter, *sample_window*, across the entire data period. Vertical spans will appear colored green or red corresponding to a win or lose respectively, with the box width matching the game duration.

3.2.2 /graph/game/{game_id}

Provided a *game id* and a target data source (Reddit, Twitter, Both), this endpoint generates a graph of the home and away team sentiment as an exponential rolling average for a window two days before and two days after the game commence time. Similar to **/graph/team/{team_name}**, the box span corresponds to the game duration, and the color represents the winning team's plotting color. This method is an extension of the Project 2 graphs which did not utilize a rolling average.

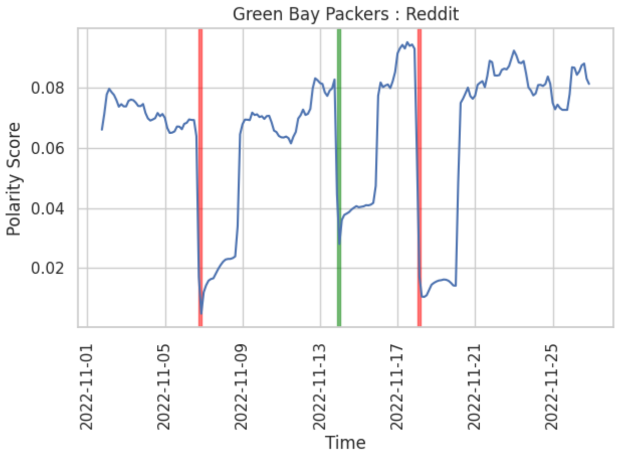


Figure 3: /graph/team/{team_name} sample response

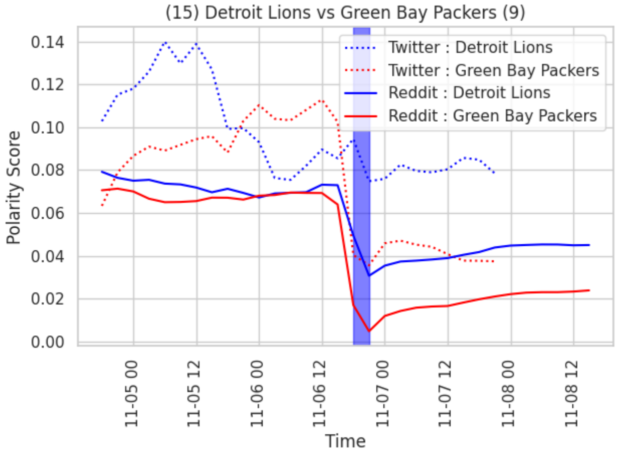


Figure 4: /graph/game/{game_id} sample response

	Detroit Lions	Green Bay Packers	Delta
Reddit	6.989	6.542	0.447
Twitter	5.256	4.463	0.793
DraftKings	+151 (40%)	-178 (64%)	329

Table 1: Aggregate Difference & Average Odds Output

3.2.3 */graph/odds*

Provided a *game id* and bookmaker, this endpoint generates a time-series graph of the betting odds leading up to the game time, which is denoted by the vertical dotted line. American head to head point odds are converted to a percentile probability per the same method used in Project 2; this method is directly ported from the Project 2 graphing.

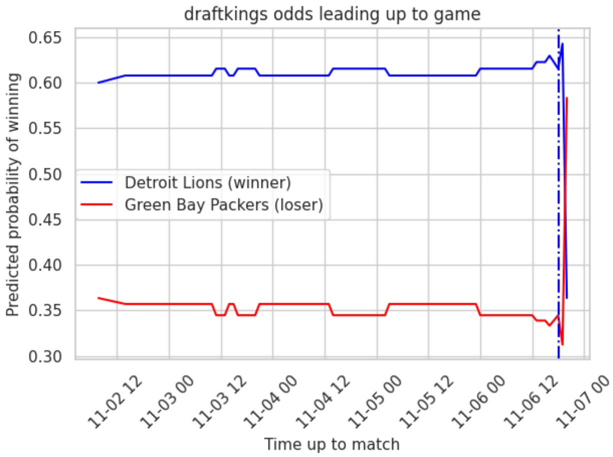


Figure 5: /graph/odds sample response

3.3 Aggregate Difference

3.3.1 */diff/{game_id}*

This endpoint is a mirror of /graph/team/{team_name}, except that it calculates a quantitative difference between the trend lines of the home and away team. An aggregate difference is calculated by computing the area under the curve up to the start of the game, exclusive, via Simpson's Rule[4] and then taking the difference. If the team with the greater sentiment wins, then it is considered *theory supporting*; aggregate difference of sentiment will be the primary mode of analysis for this report.

3.4 Sorts

3.4.1 */positivity_sort/{data_source}*

Provided a target data source (Reddit or Twitter), this endpoint returns a list of team names sorted by the aggregate positivity as defined in the aggregate difference function.

3.4.2 */most_positive/{data_source}*

This is a wrapper method of **/positivity_sort/{data_source}** that returns the *most* positive team only; for the segregated data set, Baltimore Ravens were the most positive sentiment community on average.

3.4.3 */least_positive/{data_source}*

This is a wrapper method of **/positivity_sort/{data_source}** that returns the *least* positive team only; for the segregated data set, the Las Vegas Raiders were the least positive community on average.

3.5 DataFrames

3.5.1 */df/{team_name}/{collection}/{mode}*

As an alternative to the graphing functions, the raw DataFrames used to generate the seaborn plots can be fetched as a JSON object using this endpoint given the same parameters as listed above.

3.6 Odds

3.6.1 */odds/odds_ts*

Similar to **/df/{team_name}/{collection}/{mode}**, this endpoint returns the raw DataFrame of time series data for all bookmakers for a given game id.

3.6.2 */odds/odds_avg*

This is a wrapper method of **/odds/odds_ts** that takes the average of all odds collected before the game commence period.

4 NextJS Dashboard

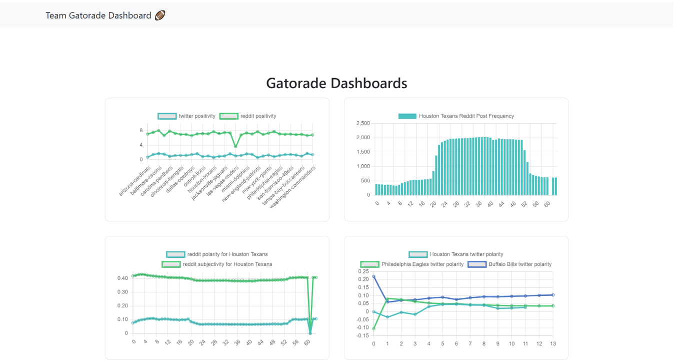

Team Gatorade Dashboard 🏈
Gatorade Dashboards
twitter positivity
reddit positivity
Houston Texans Reddit Post Frequency
reddit polarity for Houston Texans
reddit subjectivity for Houston Texans
Houston Texans twitter polarity
Philadelphia Eagles twitter polarity
Buffalo Bills twitter polarity

Figure 6: Screenshot of online dashboard for showcasing API calls

Our secondary deliverable is a front-end NextJS based analytic dashboard available at https://dashboard.ds.bxb.gg. The purpose of this dashboard is to display interactive plots that showcase calls to our API. It is a very lightweight dashboard with only a few graphs meant for demonstration purposes. The API requests are made asynchronously in NodeJS and converted to ChartJS graphs. The dashboard code allows for multiple API calls to be represented on the same figure. Filters can then be selected to choose what data is shown on the graph.

5 Rudimentary Data Analysis

5.1 Sentiment Correlation to Outcome

In our preliminary analysis, we highlighted the November 22nd, 2022 game of Arizona Cardinals vs San Fransisco 49ers as a candidate for displaying a correlation of community sentiment to game outcome. Previously the plotting did not utilize a moving average as seen in Figure 7; when comparing it to Figure 8, the separation is plainly visible. According to our **/diff/{game_id}** calculation, the San Francisco 49er held an aggregate sentiment score of 7.490, and the Arizona Cardinas held an aggregate sentiment score of 6.903, for an aggregate difference of 0.587 in favor of the San Francisco 49ers.

In 34 of 51 games (66.6%), Reddit sentiment data predicted the winner of the game correctly, and in 29 of 51 games (56.9%), Twitter sentiment data predicted the winner of the game correctly. Notably, in 11 of the 34 games it predicted correctly, but disagreed with the bookmaker odds' prediction. Betting odds were fairly homogeneous; all 22 bookmakers advertised odds that would indicate winning teams consistent with their peers. For the bookmakers, their odds predict the correct winner in 33 of the 51 games tracked (64.7%).

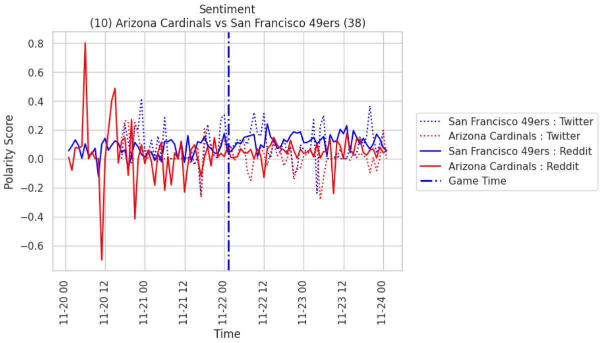


Figure 7: Project 2 Preliminary Graph

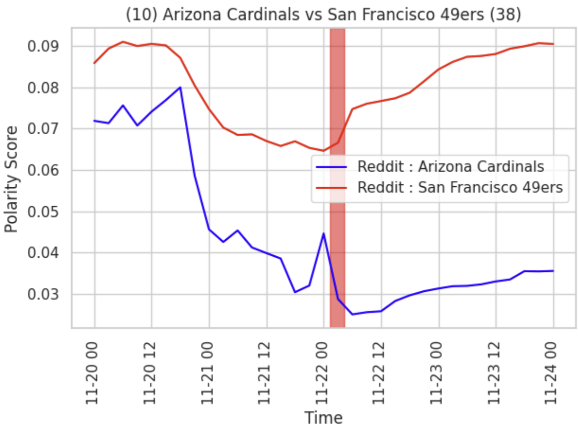


Figure 8: Figure 7 moving averages

If we assume that betting on a sports team at random has a 50/50 chance of success, and apply a Binomial Test to the sentiment data; the Reddit sentiment predictions would have a z-score of 2.24 and p = 0.013, and the Twitter sentiment predictions would have a z-score of 0.84 and p = 0.200; therefore the Reddit sentiment predictions are statistically significant and the Twitter sentiment predictions are not statistically significant. The disparity between Twitter and Reddit data is likely due to the data outage which would cause some mid-month games to have little to no Twitter data to work with. Due to time and mathematical restrictions, calculating an EV benefit from utilizing sentiment data in conjunction with, or normalizing for bookmaker odds will be omitted.

6 Conclusion

6.1 Limitations

In addition to data losses that were covered previously, our group has identified the following data limitations. First, the Twitter annotations albeit verbose may be too limiting; this is because only about

32% of tweets are said to have these annotations. Therefore, we are only collecting approximately 32% of 1% of Twitter talking about a specific team; consequently, even binning at 90 minute increments can cause buckets of only 1 tweet which shift our analysis away from "Wisdom of the Crowd" to "Wisdom of That One Fan". Similarly, our Reddit analysis is only using team subreddits; the contents of a comment is not being used to classify a team association so actions such as brigading or commentary about another team are not accounted for- likewise, this means popular shared subreddits such as **/r/nfl** are not currently used since we cannot determine the target of a given comment. Lastly, TextBlob's sentiment analysis labelled approximately a third of documents as objective neutral (0.0, 0.0); likely because there was insufficient text or the text was in a format that was difficult to parse. Currently sentiment analysis graphs integrate these data points into the average sentiment; however segregating positive and negative documents may reduce the effect of the central limit theorem and showcase trends currently unseen.

6.2 Future Works

As aforementioned in our Project 2 report, it may be beneficial to train a team classifier to allow a wider dataset, including **/r/nfl**, and then reapply this classifier to Twitter data to validate it's accuracy. Moreover, an analysis could be done across an entire sports season to get long term trends to see if consecutive wins or losses has a significant impact on sentiment, and a study could be conducted to use the predicted winners from sentiment and frequency data prior to a game are profitable signals. Moreover, a future analysis of this data, perhaps with a more complete data set that weighs bookmaker odds, or historical team performance into the ground-truth probability calculations may introduce different findings for statistical significance.

Researchers may also want to focus on the text content and sentiment during and after a game to train hate speech models or identify team rivalries/interactions. In our moving average graphs, we found a consistent dip in sentiment immediately leading into and after a game. Surprisingly, *both teams* would see a dip in sentiment; except the losing team would see a more drastic dip than the winners. It is possible that leading up and into a game that fans become nervous about their outcome, and then after the game most fans disengage from the community (potentially partying in-person) until the next game. This analysis is purely speculative and could be quantified through text content and post frequency analysis.

6.3 Conclusion

In summary, our group has found quantitative evidence that Reddit community sentiment prior to a game may be indicative of a game's outcome, however we lacked sufficient data to make the same conclusion regarding Twitter. We speculate without the data loss that Twitter may have expressed the same trend. Our computation tools and visualizations are available as OpenAPI endpoints hosted on https://api.ds.bxb.gg/docs, with a demo dashboard on https://dashboard.ds.bxb.gg/

References

[1] Phil Lord, Chris Miller, Pam Marsden, Bill Hader, and Anna Faris. 2009. *Cloudy with A chance of meatballs.* Sony Pictures Entertainment.

[2] Robert P. Schumaker, Chester S. Labedz, A. Tomasz Jarmoszko, and Leonard L. Brown. 2017. Prediction from regional angst – A study of NFL sentiment in Twitter using technical stock market charting. *Decision Support Systems* 98 (June 2017), 80–88. https://doi.org/10.1016/j.dss.2017.04.010

[3] Emma Seal, Buly A. Cardak, Matthew Nicholson, Alex Donaldson, Paul O'Halloran, Erica Randle, and Kiera Staley. 2022. The Gambling Behaviour and Attitudes to Sports Betting of Sports Fans. *Journal of Gambling Studies* 38, 4 (Feb. 2022), 1371–1403. https://doi.org/10.1007/s10899-021-10101-7

[4] Wikipedia. 2022. Simpson's rule — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Simpson's%20rule&oldid=1127617545.