CS415 Project 2 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

ABSTRACT

In this study, we aim to explore the potential use of sentiment analysis and posting frequency of social media data in predicting the expected winner of a game and identifying team rivalries and classifying reactions to game outcomes. We also seek to examine the relationship between Twitter and Reddit data in terms of sentiment and engagement, and the potential use of sentiment analysis in measuring the influence of one platform on the other in the context of sports discussions. Data was collected from noon on November 1st to midnight on November 27th, 2022 and processed using TextBlob, Cheerio, and minimal webscraping techniques. Limitations of the data due to technical failures and the limitations of TextBlob's keyword filters were addressed in the analysis.

1 INTRODUCTION

This report will act as a preliminary exploration of our dataset collected from a November 1st, 2022 at noon to November 27th, 2022 at midnight UTC, collecting from the Twitter 1% Stream, Reddit, and the Odds API. Our original goal was to substantiate the "wisdom of the crowd" phenomena by analyzing the volume and engagement sentiment towards sports teams on Reddit and Twitter against the odds and outcomes reported on sports betting sites such as FanDuel, DraftKings, and Barstool sports. By extension we also hope to be able to identify team rivalry dynamics or game outcome reactions, and identify the similarity between Reddit and Twitter communities regarding sports discussion. Due to data constraints, this analysis will be focused on teams within the NFL. Plotting was constructed using Seaborn, a wrapper of matplotlib, and TextBlob for sentiment analysis. Some additional data was fetched such as game outcomes and player rosters using webscraping techniques via Cheerio.

1.1 Supplementary Plots

The present report presents a subset of the graphs (of hundreds) generated from our analysis. For a complete set of graphs in their original resolution, including metrics on betting odds, sentiment scores, and posting frequencies for each game in our data set, please refer to the *plots* branch of our implementation repository.

1.2 Research Questions

(1) Can sentiment analysis and posting frequency of social media data be used to predict the expected winner of a game?
(2) Can sentiment analysis and posting frequency of social media data be used to identify team rivalries and classify reactions to game outcomes?
(3) What is the relationship between Twitter and Reddit data in terms of sentiment and engagement, and can sentiment analysis be used to measure the influence of one platform on the other in the context of sports discussions?

1.3 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[1] 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[2] 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 pipeline experienced a catastrophic failure due to misconfigured restart logic for network failures. To be brief, our use of the process manager (PM2) to initiate restarts on crashes failed to account for asyncio event loop errors; as such, the inner event loop monitoring the Twitter stream failed without alerting PM2. This issue went unnoticed from November 8th until our Project 2 proposal feedback meeting on November 14th, when the process was manually restarted. A secondary outage occurred on November 18th and 19th, after which a permanent fix was implemented. Despite the missing data, preliminary analysis was able to be conducted successfully. 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.

2.2 Data Preparation

To prepare our dataset for analysis, we used a combination of TextBlob, Cheerio, and webscraping. TextBlob was used to compute sentiment polarity and subjective scores, and was multiprocessed for computational speed-up. TextBlob calculates sentiment polarity and subjectivity scores based on the positivity or negativity, and subjectivity or objectivity of a given text. These scores range from -1 to 1 and 0 to 1, respectively, with higher values indicating a more positive or subjective sentiment. Unfortunately, due to the limitations of TextBlob, approximately 35% of the documents could not be classified positive or negative, and were assigned a strictly neutral sentiment (0.0, 0.0). This may be due to the fact that these records contain insufficient data for TextBlob's keyword filters to accurately classify them.

Cheerio was used to extract NFL player names from footballdb.com, and FuseJS was used for a fuzzy search of player names against Twitter's context annotation CSV to create a mapping of team rosters to entitlement ids. This mapping allowed us to associate tweets with a given team, and in the future may be used to train a classifier for Reddit. Similarly, NFL game times and scores were webscraped from pro-football-reference.com and correlated with the betting game ids used by the Odds API.

2.3 Data Classification

Source	Record Count	Size
Reddit Comments (All Sports)	11.7M	5.38GB
Reddit Comments (NFL)	1.3M	578.9MB
Reddit Submissions (All Sports)	181.4K	190.93MB
Reddit Submissions (NFL)	15.9K	16.7MB
Twitter (Raw)	58.9M	5.95GB
Twitter (All Sports)	2.0M	8.46GB
Twitter (NFL)	71.6K	318.4MB
Odds API (All Sports)	15.3K	299.4MB
Odds API (NFL)	1.4K	27.8MB

Figure 1: Collected Data Counts

Figure 1 shows the number of records collected per data source and the dimensions of each collection when considered as either all sports or NFL data. It also shows the storage consumption of each collection. Interestingly, the collection of 2 million Twitter documents consumes 8.46GB, while the collection of 11.7 million Reddit documents only uses 5.38GB. This likely stems from the abundance of context annotation metadata associated with the Twitter data, which is not present in the Reddit data.

Figure 2 denotes the rate of all ingested data; Figure 3 is a refactored version of Figure 2 that only pertains to plots with the post-filtered, NFL only data. From these plots, it is apparent outside of game-day, subreddits tend to be dormant whereas Twitter sports discussion tends to be more constant, albeit lesser than Reddit's game-day peaks.

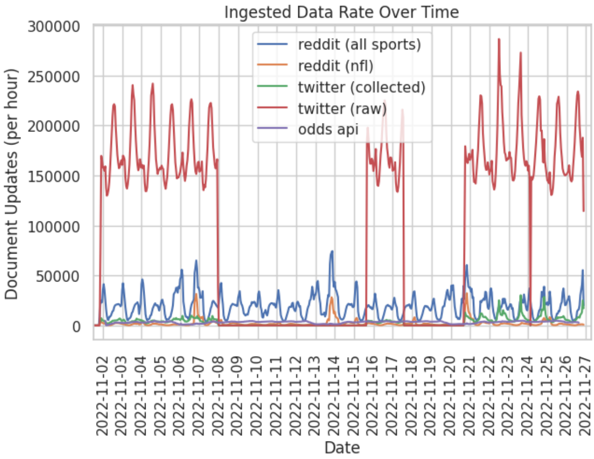

Figure 2: Rate of ingested data without filtering

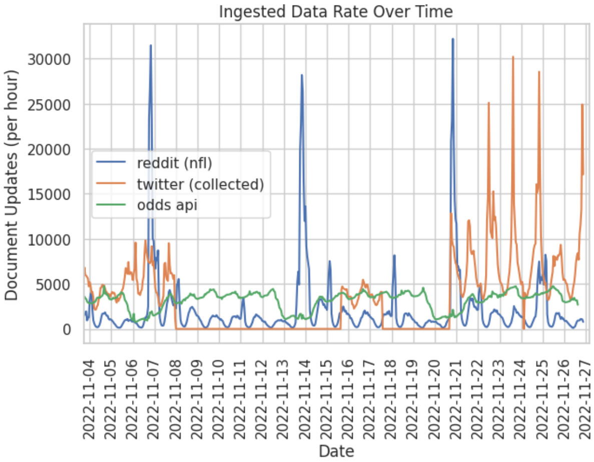

Figure 3: Rate of ingested data, NFL data only (Figure 2 zoomed in)

In our study, we utilized the Odds API to gather betting data from 22 sports betting sites, as shown in Table 5. Specifically for the NFL, we captured data for up to 71 games with approximately 6-8 timestamps per game. We plotted the odds leading up to the games reported by the three largest sports betting sites (FanDuel, DraftKings, and Barstool), as denoted in Figure 4. Surprisingly, not all NFL games are supported by every bookmaker, with some bookmakers consistently at 57 tracked games instead of 71; this likely is caused by future games which are not listed yet.

The odds were converted to probabilities using the following equation:

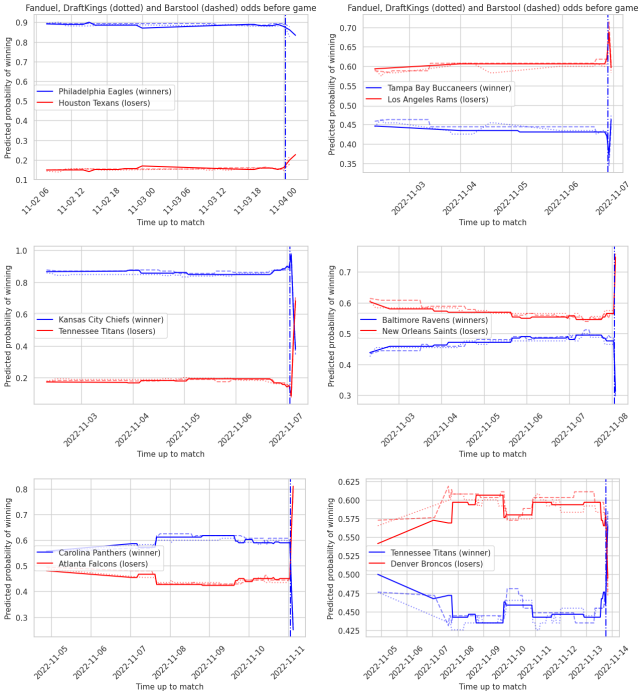

Figure 4: The odds on sports betting sites leading up to the game and during the game

Bookmaker	NFL Games	# of Games	Size
Barstool*	71	699	17.37MB
Betfair	57	494	8.05MB
BetMGM	57	691	8.53MB
BetOnline	65	762	15.95MB
BetRivers	71	702	18.56MB
BetUS	57	679	13.61MB
Bovada	68	633	11.61MB
Circa Sports	57	461	14.48MB
DraftKings*	69	695	17.95MB
FanDuel*	71	806	19.09MB
FOX Bet	57	641	8.07MB
GTbets	69	749	9.99MB
Intertops	55	503	4.57MB
LowVig	65	772	15.87MB
MyBookie	69	791	15.96MB
PointsBet	71	776	15.75MB
SugarHouse	71	698	18.34MB
SuperBook	66	780	9.34MB
TwinSpires	71	698	8.32MB
Unibet	71	697	18.84MB
WIlliam Hill	57	785	15.75MB
WynnBet	57	783	13.44MB
Totals	N/A	15,295	299.44MB

Figure 5: Odds Collection Data Counts

$$probability(x) = \begin{cases} \frac{100}{x+100} & x > 0 \\ \frac{-x}{-x+100} & x < 0 \end{cases} \tag{1}$$

where x represents the odds of each team on the sports betting site. The resulting graphs indicate that the main sports betting sites generally follow similar odds, as expected. This suggests that arbitrage betting is unlikely to be successful on these main sites. Furthermore, the odds remain relatively stable leading up to the games, with rare exceptions.

3 REQUIRED PLOTS

Per project 2's implementation specifications, graph of the 1% Twitter stream post frequency has been provided in Figure 6; due to the data outages, we have expanded the graph to contain up to November 27th, 2022. Similarly, Figure 7 provides a graph of $r/politics$ submissions from November 4th, 2022 to November 14th, 2022 (inclusive).

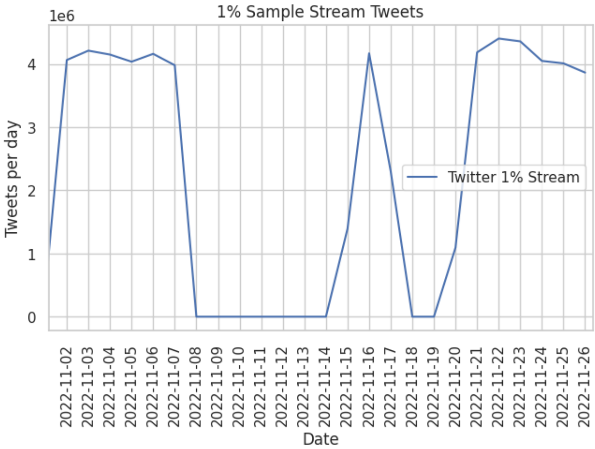

Figure 6: Twitter 1% Stream (daily)

As aforementioned, our Twitter data collection pipeline experienced an outage from November 8th-14th, and November 18th and 19th, which can be seen in Figure 6. Referencing Figure 9, which bins the data hourly, may provide better inside into the circadian mode of the posting frequency.

Regarding Figure 7, the most active time periods shown are November 7th, 9th, and the 13th. These results make sense in light of the United States midterm elections took place November 8th. For November 13th we believe the massive spike to be due in part to the news of Donald Trump's announcement to seek reelection.

4 DATASET ANALYSIS

4.1 Data Ingested Rates

In Figure 8, we see the total Reddit data, split up between all sports and NFL. As we see, similar to Figure 9, there are the midday and weekend modes. Going deeper, days that feature a lot of sporting events feature a lot of activity. As sports like NBA, MLB, and NHL have well spread-out schedules that do not feature any one day too heavily, those sports do not lead to any noticeable spikes as

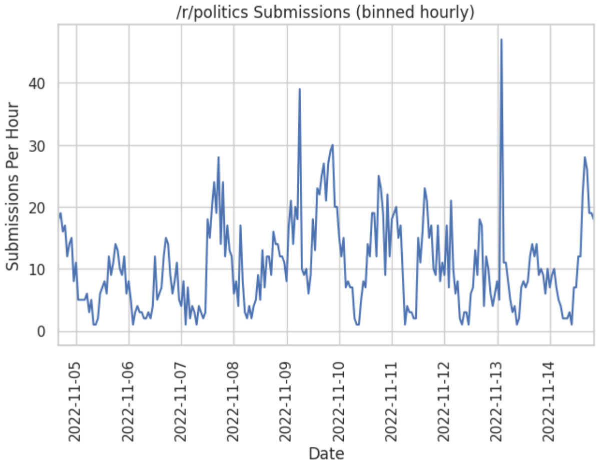

Figure 7: r/politics Submissions (hourly)

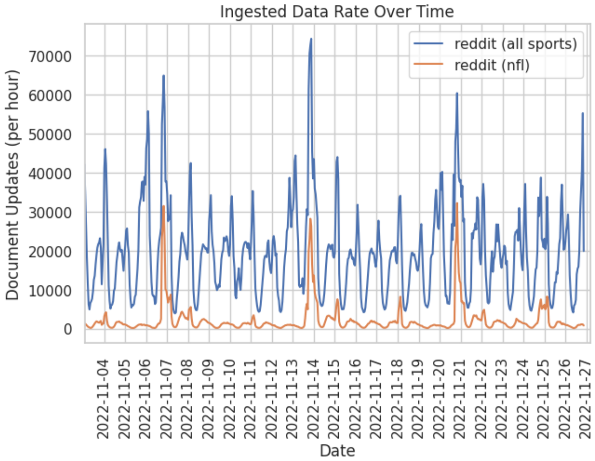

Figure 8: Reddit's data ingest rate, all sports vs NFL teams only

compared to the NFL and NCAA Football. The NFL features games primarily on Sunday, as well as singular games on Monday and Thursday. These spikes show up heavily on the nfl-only line, but also show influence over the total Reddit data collected. With NCAA Football, their games are mostly scheduled on Saturdays, which can explain the peaks on Saturdays in this data.

4.2 Rivalry Preliminary Analysis

Figure 10 and Figure 11 demonstrate highly reactive sentiment responses to game wins and losses. Prior to the outcome of the game, the sentiment of both teams appears to be similar, but following the game, a clear divide is evident, with the winning team displaying overwhelmingly positive sentiment and the losing team displaying negative sentiment. It is possible that a correlation exists between fan sentiment and the outcome of the game, as the largest spikes in sentiment prior to the game appear to coincide with the winning team.

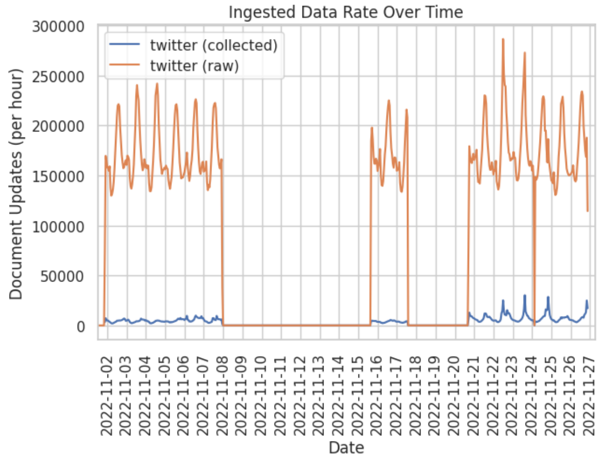

Figure 9: Amount of Tweets encountered and collected, binned hourly

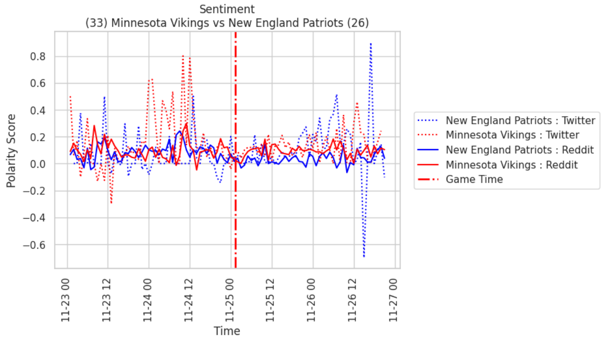

Figure 10: Example 1 of reactive sentiment following game outcome

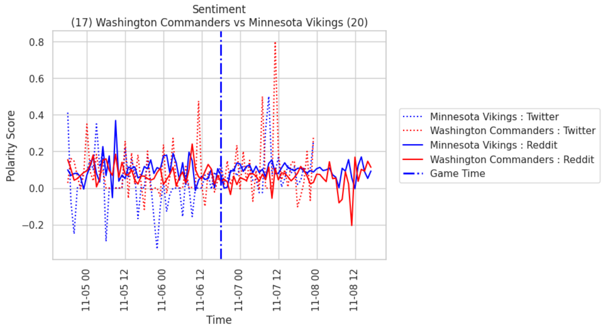

Figure 11: Example 2 of reactive sentiment following game outcome

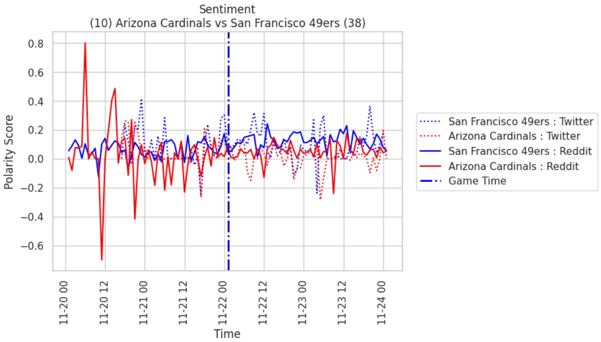

Figure 12: Example of sentiment correlating with game outcome

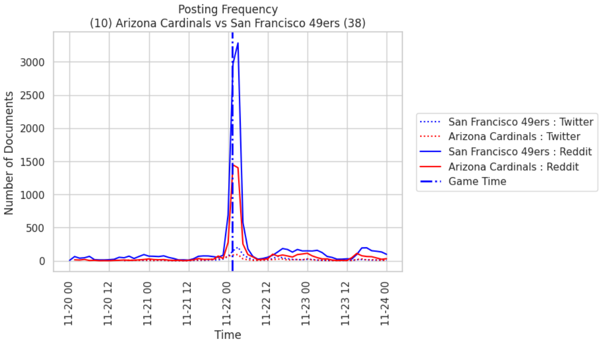

Figure 13: Posting frequency of Figure 12, binned every 90 minutes

4.3 Sentiment Correlation to Outcome

A potential example of a relationship between fan sentiment and game outcomes is seen in the Arizona Cardinals vs San Francisco 49ers game on November 22nd, 2022, as illustrated in Figure 12. The 49ers are consistently more positive than the Cardinals leading up to the game, and the Cardinals experience a significant negative event on November 20th, resulting in an average sentiment polarity of -0.6 over a 90-minute period. That said, the extreme variance on November 20th, may be caused by insufficent data. Figure 13 highlights extreme inactivity in both team's subreddits on non-game-days, meaning the sentiment scores may be influence by only a handful of comments in total; further investigation into this relationship will be necessary in Project 3.

4.4 Betting Odds Update Frequency

In Figure 14, the gray windows highlight Sunday, and dashed red lines correspond to game commence times. Bookmakers typically avoid updating their odds on game-day, which may be because the variables are the least volatile and no longer require updating, or because they fear that doing so may annoy customers.

Despite this lull, Figure 4 demonstrate NFL game head to head betting odds appear to be fairly consistent leading up to a game; when considered in conjunction with Figure 14 this indicates only minor adjustments are mades to the odds and the largest established betting sites may not be ideal targets for arbitrage betting via sentiment shifts. That said, in Figure 15 exception may be made for the Tennessee Titans vs Denver Broncos game on November 13th which saw erratic odds adjustments leading up to the game, which were inconsistent between the three sports book makers.

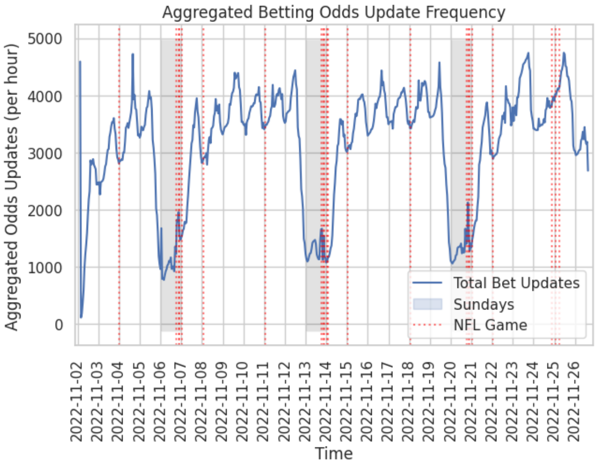

Figure 14: All sports odds update frequency graph; showcases reduction of updates on Sundays

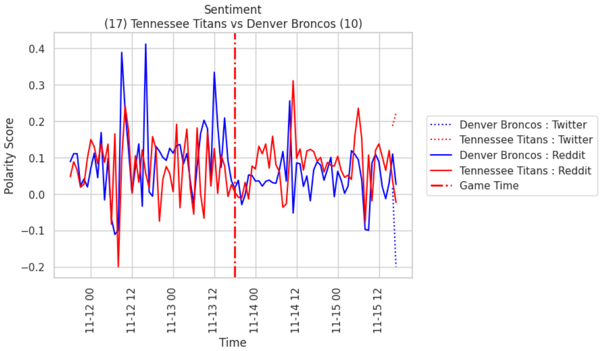

Figure 15: Counter example to sentiment analysis outcome correlation.

4.5 Social Media Sentiments

Comparing general Twitter and Reddit sentiment analysis, we see, potentially surprisingly to some, positive-leaning sentiment. In Figure 16 and Figure 17, both charts stay between .05-.1 for polarity, which indicates slightly positive sentiment amongst the data. However, Reddit and Twitter have differing subjectivity ratings. Reddit's subjectivity remains fairly stable at around .4. Twitter on the other hand remains between .1 and .2, and noteably has a lower subjectivity in the last week of the data compared to the first.

This difference in subjectivity most likely stems from the more public nature of Twitter compared to Reddit. On Reddit, the people viewing your posts will more likely be like-minded individuals

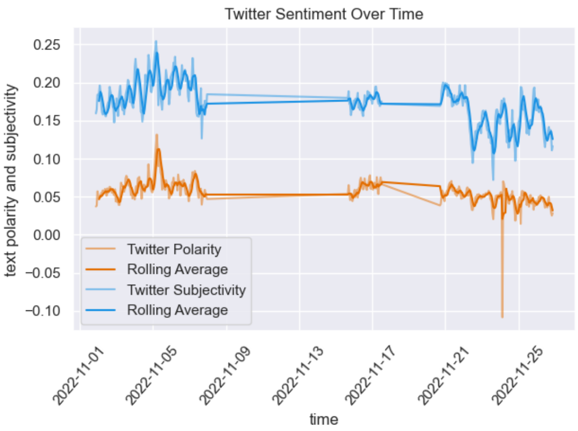

Figure 16: Twitter sentiment over time

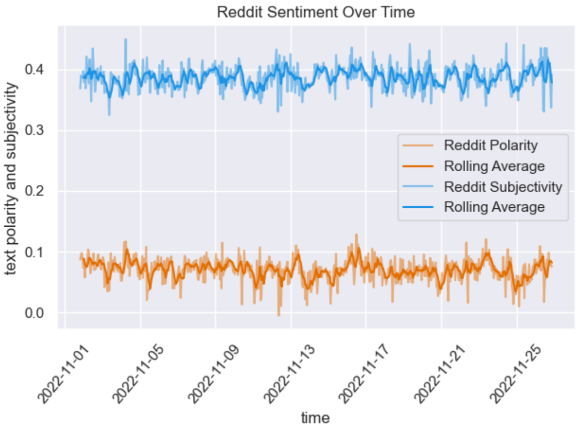

Figure 17: Reddit sentiment over time

as you post in your teams subreddit. This like-mindedness, and subsequent lack of judgement, can lead to an increase in subjectivity. However, with Twitter, the more public nature can lead to less subjectivity in posts as your tweets are exposed to a larger, more diverse audience.

5 CONCLUSION

5.1 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.

5.2 Future Works

In the future, 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.

5.3 Conclusion

In summary, our group has found qualitative evidence that community sentiment prior to a game may be indicative of a game's outcome. Following our preliminary research, our potential research questions going into Project 3 are: "Can sentiment analysis and posting frequency of social media data be used to predict the expected winner of a game?", "Can sentiment analysis and posting frequency of social media data be used to identify team rivalries and classify reactions to game outcomes?", and "What is the relationship between Twitter and Reddit data in terms of sentiment and engagement, and can sentiment analysis be used to measure the influence of one platform on the other in the context of sports discussions?".

REFERENCES

[1] 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

[2] 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