Parlays for Days: Project 2 Proposal

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 Power Rank: The Football Analytics Resource Guide

1 INTRODUCTION

Our goal is to substantiate the "wisdom of the crowd" phenomena by analyzing the volume and engagement sentiment towards sport teams on Reddit and Twitter against the odds and outcomes reported on sports betting sites such as FanDuel and DraftKings. By extension, when visualizing our data set will be able to identify team rivalry dynamics and overall team popularity in Twitter and Reddit communities, and classify the toxicity of sport communities.

2 PROPOSED EXPERIMENTS

2.1 Wisdom of the Crowd

To evaluate the relationship of the "Wisdom of the Crowd", we currently track the total volume of content relating to each team, and calculate an average sentiment towards each team over time, with game times and outcomes. To classify Twitter data, we likely will rely on the entity names, and create aggregated "bins" of sports team members/mascots entities to identify the target of a tweet. To generate these lists, we will likely pull data from official sources such as NFL.com, and utilize the team rosters in a fuzzy search against the labeled entity names provided by Twitter's annotations. For Reddit, team subreddits could be assumed self focused, i.e, assume /r/BuffaloBills is talking about Buffalo Bills, but this could lead to false positives and wouldn't be applicable for mixed subreddits such as /r/NFL. As such, training a classifier using the Twitter entity names may be investigated as an alternative method, or metadata such as author flairs could be used to better gauge the intent and bias of a Reddit submission or comment. From both data sources, we hypothesize that a change in sentiment or volume of tweets may indicate a significant community opinion shift, and could be used to dictate the expected winner. By relying on the rate of change rather than raw output, we should be able to normalize for teams that are naturally more popular or hated than others, and only breaks from the norm would be considered statistically significant.

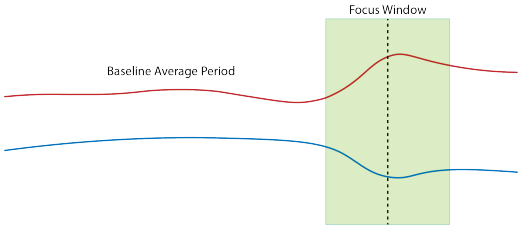

Figure 2: Example of easily detectable game sentiment disparity leading up to a game.

Take a hypothetical example in Figure 2 where, when compared to their baseline averages, Team Red has an increased sentiment leading up to their game, and Team Blue has a decreased sentiment leading up to the game; from our hypothesis we assume that Team Red is expected to win by the community, and this could be validated against the game outcome. Moreover, this could be compared to the sports book odds to identify outliers or potential underdog victories.

Moreover, this analysis would easily translate to the analysis of bookmakers. Instead of looking at game outcomes, we could consider the betting odds, and answer questions such as "Is sentiment

correlated with the bookmakers odds, and when sentiments shift, do bookmakers odds also shift?".

2.2 Rivalries and Reactions

In addition, this visualization could see if sentiment leading up to a game is indicative of team performance. We could visualize if a fandom is a sore loser by maintaining a low sentiment score after a loss, or if certain rivalries are visually present through low sentiment or high toxicity scores leading up to a game. This analysis largely would make use of the same computational and analysis methods as section 2.1, but focused within a different context.

2.3 Cross Impact of Twitter and Reddit Data

We will compare and contrast our analysis between Twitter and Reddit to see if there are any relationships in the data. Notably for Twitter's influence on Reddit, we can utilize the percentage of submissions linking to Twitter and the amount of engagement these submissions receive. Likewise, we will analyze if the sentiments match across platforms or if the platforms have a noticeable affect on each other in regards to sentiment. Monitoring Reddit's influence on Twitter will be more difficult due to the lack of direct links, but we can search for mentions of subreddits and do a keyword prevalence comparison between the two platforms. We can also run sentiment analysis on both platforms to gauge the amount of influence present in these specific kinds of posts. We can also keep track of the number of posts to examine the influence these posts have on overall sentiment on sports related posts.

The algorithms we will utilize include basic graphing, machine learning sentiment detection, and keyword detection. The sentiment model will most likely be a PyTorch based pretrained model or a model from the spaCy library; spaCy will also be able to handle text blob tokenization and lemmatization. We will also most likely only be analyzing English posts as that is what sentiment analysis models are trained on and what we would have a better time of understanding the results of.

3 DATASET SCOPING

At this time we plan to focus our analysis on football (NFL and/or college football); we do have minimal data on the MLB post-season, but it is unlikely to be sufficient for in-depth analysis. To streamline the NLP data pipeline and improve sentiment analysis accuracy, we will also limit our dataset to tweets identified as English, and English speaking subreddits. This decision is deemed acceptable due to the fandom demographics surrounding the NFL and MLB as highly American, and is reinforced by the prevalance of English speakers in our collected Twitter data, as seen in Figure 3.

4 ADDITIONAL DATA

To achieve our goals, we need team roster data to inform our post to team classification models; game outcomes and scores to quantify the accuracy of social media sentiment and bookmaker odds to game outcomes; and other meta information such as player injury reports or major news articles that could explain an upset victory or external factors that influenced a game's outcome. These data points should be accessible through sources such as NFL.com, espn.com or teams sports pages and will likely be aggregated without use

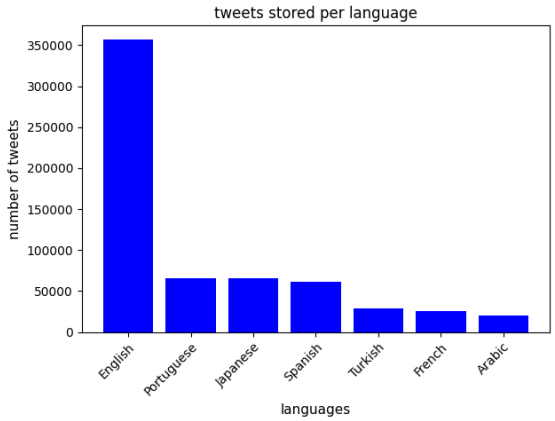

Figure 3: The number of total and stored tweets per language

of a dedicated API. We will also need sentiment analysis scores which likely be processed using spaCy and TextBlob, or through Google's Cloud Natural Language Processing model. For the latter, we understand it is a paid API, but we already are leveraging GCP 90-day promotional credits for an emergency provisioned VM in Project 1, so the web requests would be "free". Currently the biggest bottleneck in this method would be the 800,000 requests/day limit; which outside of last minute processing, should be fine for the size of our dataset; in an emergency we may opt to sample our data to reduce the total amount of comp.

5 CONCLUSION

In summary, we hope to cross reference sports community engagement and sentiment against sports betting odds and outcomes to substantiate the "wisdom of the crowd" phenomena, in addition to tracking the popularity and type of interaction between sport team communities. This will be achieved through measuring the volume and sentiment of tweets and Reddit posts against their historical averages, and analyzing their correlation to bookmaker odds and game outcomes. We will also use this traffic and sentiment data to potentially identify team rivalries and the fandom reaction to wins and losses. Lastly, we will analyze the overlap of content shared between Reddit and Twitter communities through the use of keyword and URL analysis.