Parlays for Days: 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: Stonks

1 INTRODUCTION

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

Our group recognizes sports teams inherently attract different fan base sizes; rather than relying on raw volume metrics, we plan to normalize our analysis by focusing on the sentiment and volume trends changes approaching a game. We hypothesize that optimistic fan bases will post more often leading up to a game and in corollary, pessimistic teams will engage less often leading up to a game- by using this delta we can predict the community's favored winner, and identify optimal bets when the community expectation is inverse that of the betting odds. As an additional point of analysis, we may cross reference the community sentiment against a team's or player's rated strength to see if they are correlated, or use this as a weighting factor.

We will strive to design an agnostic pipeline, enabling the concurrent tracking of multiple large sports, such as the NFL, MLB, NBA, MMA, UFC, etc, simultaneous, but due to time, data set, or hardware limitations, we may opt to narrow the project scope to only a single organization, such as the NFL at the time of implementation.

The results of our work will quantify the communities' "wisdom of the crowd" ability to predict the outcome of sports games or form optimal bets, and track how the outcome of sports games affect a communities sentiment internally and towards other sports communities.

2 DESCRIPTION OF DATA SETS

2.1 Reddit Data Set

As a preliminary data Reddit set, we referenced the /r/ListOfSubreddits directory for a list of sports and sports team subreddits; this list contained 218 entries, which include mainstream sports, niche sports,

such as /r/curling or /r/racquetball, and the associated franchise team subreddits, such as /r/buffalobills or /r/bostonceltics. Moreover, a breadth first search was completed aggregated the subreddits mentioned in the sidebars for a depth of 2, resulting in 4,417 results. Then, the group was filtered into a list of subreddits with more than 5 comments per day or 1 post per day (based on the last 90 days averages of Pushshift data) and was categorized as ad_category of "Sports" by Reddit; resulting in 232 qualifying subreddits.

The set of 232 subreddits identified here will be used for future data estimates, but it is likely our scope will narrow to track a specific subset of sports.

2.2 Twitter Data Set

The bulk of our data analysis will be utilizing the Twitter V2 Volume Stream. We acknowledge retaining the unfiltered data stream is unsustainable, so metrics such as relevant hashtags and keywords relating to the sports or names of players associated with the sports we are tracking will be used to refine set of tweets.

Additionally, if the 1% Volume Stream is insufficient, we may opt to leverage our approved access to the Academic Researcher 10M tweet/month rate limits and utilize the Twitter V2 Filtered Stream for data matching our previously defined filter rules.

2.3 Sports Betting Data Set

For our third data source, we will collect sports betting odds from the bookmakers, DraftKings and FanDuel; both services lack official API documentation, but community compiled unofficial documentation exists that references unauthenticated internal API calls, we have not investigated the enforced rate limits, but it is unlikely we would need to poll DraftKings more often than an enthusiast human user would be expected.

Alternatively, the Odds API is a freemium API that aggregates the listings of all major betting markets (in a way similar to Pushshift for Reddit), but has a free tier capped at only 500 requests per month, with paid plans starting at $20.00 for 20,000 requests per month; this likely will be the most streamlined option, albeit paywalled. Lastly, if the preceding options fail, we can defer to in-browser webscraping through tools such as Selenium or BeautifulSoup.

3 ESTIMATED RESOURCE CONSUMPTION ["NAPKIN MATH"]

3.1 Twitter

On September 26th, 2022, we sampled the Twitter V2 Volume Stream for approximately 24 hours storing all tweets received in a MongoDB collection. During this collection period, we tracked 3.8M tweets with an average of 44 tweets per second, and stored 173.4MB of compressed data on disk, 694.1MB uncompressed. A case-insensitive dumb text search for "nfl" retrieved 6,342 records or 0.17% of the data stream; do note the text was *not* tokenized, so false positives such as cornflake, would be included in this figure.

Assuming this metric holds for other keywords, and our group tracks 60 keywords with zero overlap (highly unlikely), this would correlate with 10% of the Twitter V2 Volume Stream, corresponding to a range of 17MB - 69MB of data collected per day, dependent on compression rates. For simplification, we will assume similar compression and use 20MB/day going- but with the previous assumptions, this figure *will be a liberal overestimate*. Extrapolating to 47 days, results in 1.4GB of data usage.

Table 1: Summary of Expected Storage Consumption

Data Set	Storage Consumption (MB)	Total Usage (GB)
Twitter	20 MB/day	940 MB
Reddit	1,000 MB/day	47 GB
Sports Betting	325 KB/day	15.3 MB
Total	1020.3 MB/day	73.4 GB

3.2 Reddit

To quantify our qualifying 232 subreddits, we fetched the submission and post counts for the last 90 days via the Pushshift API and in total, this data set attributes to 6.7M submissions and 293M comments. According to a /r/dataisbeautiful submission, /r/nfl averages 180 characters per comment, and /r/nba averages 120 characters per comment. Accounting for title, usernames, and other metadata, we will assume 300 bytes per document, resulting in 1,000MB/day or 47.0GB when extrapolated to the last day of classes. We can drastically reduce this figure by focusing on an individual sport instead of a broad spectrum of sports.

3.3 Sports Betting

Lastly, we expect the sports betting endpoints to contribute the least amount of data from our generated data sets. Assuming usage of the lowest paid tier of the Odds API, or a reasonable polling rate, we could expect to fetch up to 650 bets odds/outcomes per day. Over the span of 47 days this would provide us 31K bet odds/outcomes. Accounting for team names, timestamps, odds, and other meta data, each document could be up to around 500 bytes, resulting in 0.325MB/day or 15.3MB by the end of the data collection period.

3.4 Summary

In total, with liberal figures, and no limited scope to the sports followed, we expect to utilize 50GB between October 23rd, 2022, and December 9th, 2022 in data storage. By large, this is an intentionally high estimate with uncertain values biased towards being the upper threshold of expectations.

4 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 done through the measuring the volume of tweets and/or Reddit posts against their historical averages, and how they may correlate to potential betting odds of games; keyword and hashtag recognition of prominent players or sports team names, and the potential weighting against traditional sports odds creation such as player and team performance statistics.