Parlays for Days: Project 1 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 Sports Geek: Common Football Bets

1 INTRODUCTION

We have successfully implemented an asynchronous data stream using Python 3.11 and MongoDB against the Twitter V2 Volume Stream, Reddit JSON API and the Odds API. Our tooling heavily relies on the packages, *aiohttp*, *asyncio*, and *motor* to ensure non-blocking operations and error tolerance. Each data source is run as a unique process and managed by *PM*2 for automatic restarts and job scheduling. We have deferred any sentiment/NLP analysis and instead have opted to rely on meta heuristics or existing data classifications to filter our data. We have also deferred the narrowing of scope to a specific sport or set of sports, and are in good footing to generate comparisons of community sentiment to the predictions and outcomes of 22 US bookmaking entities in the United States in Project Two.

Due to technical difficulties with the hosted environment, our stable data period starts at approximately November 1st, 2022, 6PM ET, but all data is currently hosted on the university provided virtual machines.

2 IMPLEMENTATION

2.1 Reddit Data Set

For our Reddit implementation, we maintain the subset of 253 subreddits (of originally 4,471) aggregated via the /r/ListOfSubreddits page of sports and sport team subreddits and a DFS search of subreddits mentioned in a community's wiki/descriptions, as outlined in our proposal paper. This figure includes all seed subreddits (those originating from /r/ListOfSubreddits' list), plus all discovered subreddits with *ad_category* set to *sports*. As such this represents a broad spectrum of all major sports (NFL, NBA, MLB, etc), in addition to their associated team subreddits (Buffalo Bills, Golden State Warriors, New York Yankees, etc.). Potential such as */r/barefoot* satisfy these constraints, but their performance impact and ease at which to apply filters in future stages makes this broad net of subreddits deemed acceptable.

For data collection, we opted to leverage multireddits via the JSON reddit endpoints to fetch in pages of 100 posts; tracking both submissions (*/new.json*) and comments (*/comments.json*); we are also using the password grant OAuth2 flow for the increased rate-limit. A bounded-set is used to keep track and yield new content by comparing content ids with the previous page; rebalancing events are triggered if the number of duplicates fall below a threshold of 15.

To prevent data loss, naive forms of load balancing have been implemented to detect and move popular subreddits into lower traffic, or multiple search bins via greedy multiway number partitioning. At the current scope of subreddits, often only one bin per post type is necessary, but at peak hours (such as during an

NFL game), the rebalancing event may trigger. Rebalancing logic is handled in memory and frequency statistics are saved to MongoDB every cycle (all subreddits processed once) as a exponential moving average to improve restart performance.

2.2 Twitter Data Set

When processing the Twitter V2 Volume Stream our group heavily relies on Twitter's native Tweet annotations system to filter irrelevant data. Twitter annotations consist of *domains* and *entities*. Domains are general categories such as "Sports Team", or "Journalist", and entities are human-filtered keywords that are placed into domain categories by Twitter appointed domain experts. Entities can belong to multiple domains, and domains can contain multiple of the same entities. We are currently tracking 17 sports related domains, which encapsulate 22,044 of the 144,753 labelled entities in Twitter's repository. Moreover, we are manually are including 312 entities that fall outside the 17 domain set but were still deemed pertinent- about half of which are $< Team >$ $Stats$ entities that are classified under domain 131, $UnifiedTwitterTaxonomy$, whereas others include bookmaker names such as "DraftKings" which falls under 47, *Brand*, both of which are far too broad to be used globally in our filter.

In terms of robustness in handling errors and timeouts, we are compliant with Twitter's recommendations, such as utilizing the new-line heartbeat to detect as disconnect if no heartbeat was heard within 20 seconds; in addition to following guidelines for retry-logic on network or rate limiting errors. Furthermore, in the event of a hard reset, we leverage our Academic Access bearer token to use the $backfill_minutes$ to minimize data loss for up to 5 minutes of downtime.

2.3 The Odds API

Usage of the Odds API is fairly straightforward and will make up a negligible amount of data storage in comparison to Reddit or Twitter. Currently we are polling the Odds API hourly for odds updates and if a discrepancy is found, a time series object is created within the game record to track the shift in odds over time. For this project we are leveraging use of multiple free accounts (of which is partially automated via 2Captcha solving + Gmail inbox reading) and have created logic to automatically cycle keys when a free token exhausts it's rate limit. In our request we need only to define the sport and type of bets, and the Odds-API can automatically aggregates up to 22 US sports betting bookers live-odds in it's response for most upcoming games. For ease of data access, we have opted to store each booker as it's own collection as it's likely we will be comparing users to a given bookmaker's odds, rather than bookmaker's odds against each other.

3 PRELIMINARY DATA EXPLORATION

3.1 Twitter

The Twitter data contains little to no false positives as determined by random searches and initial exploration of the data. Our pipeline encountered 138,000 tweets per hour and stores 2,900 tweets per hour on average with a maximum of 4,426 and a minimum of 450. This equates to an average of 2.1% of tweets that we are grabbing from the 1% stream. We graph these metrics in Figure 2a which show the total tweets, stored tweets, and percentage of tweets stored. We also graph the amount of tweets per top 7 languages (Fig. 2b) and tweets related to two popular sports NFL and MBA, versus the language of the tweet (Fig. 2c).

Anecdotally, we have found the Tweet Annotations to be fairly accurate, and can observe the correct annotations that are not a direct-text match for an entity, rather *context aware*. For example, a tweet stating "SEAHAWKS! SEAHAWKS! SEAHAWKS!" being context annotated as "NFL" demonstrates a sophisticated annotation system that we are confident in.

Our preliminary data exploration shows promising results with the amount and accuracy of the data we are storing. We see that English is the largely predominant language of football sports related tweets while it is also the predominant language of all tweets. This highlights that English tweets should be our main area of focus going forward.

3.2 Reddit

Our group recorded hourly statistics and represent them in a histogram as shown in Fig. 2d; due to technical difficulties surrounding unresponsive VMs the data window is brief. In addition to these plots, we found 15% of our initial submissions data reference a Twitter post, 38% are self posts, and 63% of submissions include a self-text. As expected the currently among the most popular subreddits are $/r/nba$, $/r/nfl$, and content activity is diurnal, following the United State sleeping pattern.

3.3 The Odds API

The initial data exploration of our sports-betting database is limited. We are currently making hourly requests for head-to-head, spreads, and game totals against 22 US bookmakers, including DraftKings, FanDuel, BarStool, BetUS, etc. We've found betting odds to be relatively static but experience shifts approaching game commencement times and may investigate increasing the polling rate closer to popular game commencement times. Currently, most US bookmakers have up to approximately 80 games available for betting, with each game having usually about 12 bets per game.

3.4 Summary

In total, our preliminary exploration of the data shows promising prospective results with the amount of relevant data we are collecting. So far we have had no data connectivity or storage issues (barring a VM that wouldn't reboot) and we hope to keep it this way going forward with our use of PM2 and MongoDB. Going forward we will do more cross analysis between the platforms we are analysing. Once we have more Odds API data we can also start creating graphs of social media posts versus the outcomes of games and eventually analyze the posts using sentiment analysis.

4 SCHEMAS AND SPECIFCS

For sake of brevity, (and the inability to figure out how to fit 200+ item tables into a LaTeX document. Refer to the implementation GitHub repository README for lists of the filtered domain list, subreddit list, and collection schemas. Similarly, for a list of whitelisted entities (excluding *<team> stats* entities which are matched by name), consult the *custom_eval.py* function source code found here.

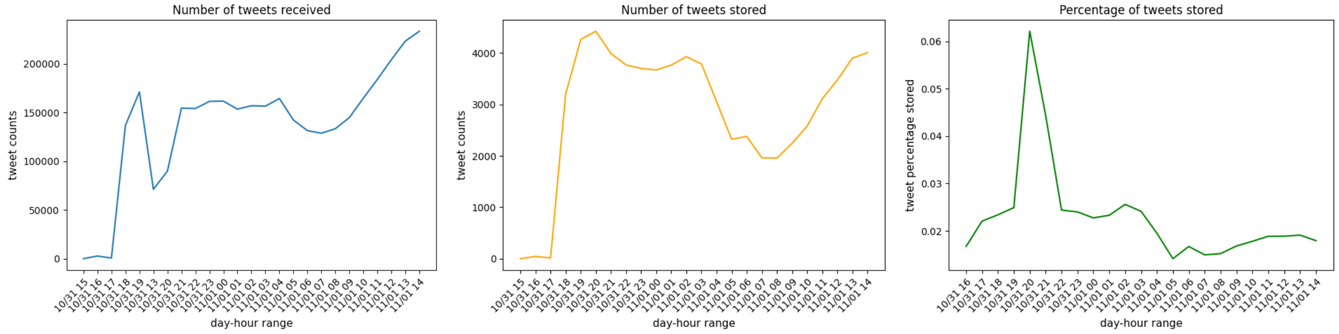
These values are hard-coded defaults that are repopulated via a database, so we can easily enable/disable more or less entities/context domains or subreddits.

5 CONCLUSION

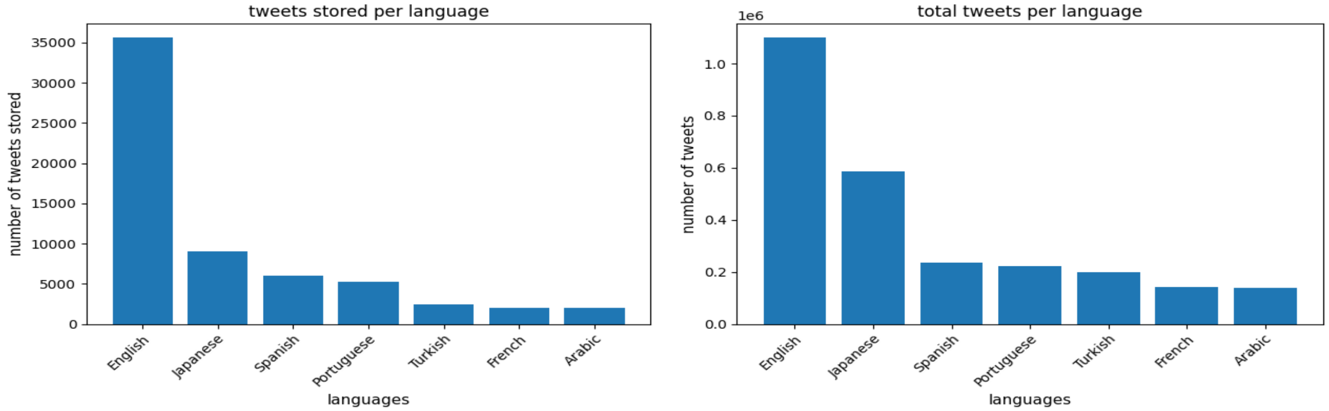
In summary, we hope to cross reference sports community engagement and sentiment against sports betting odds and outcomes to substantiate the "wisdom of the crowd phenomena", in addition to tracking the popularity and type of interaction between sport team communities. This will be done through the measuring 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.

Our implementation is asynchronous and has protections in place against hitting rate limits or inadvertently losing data when popular events take place.

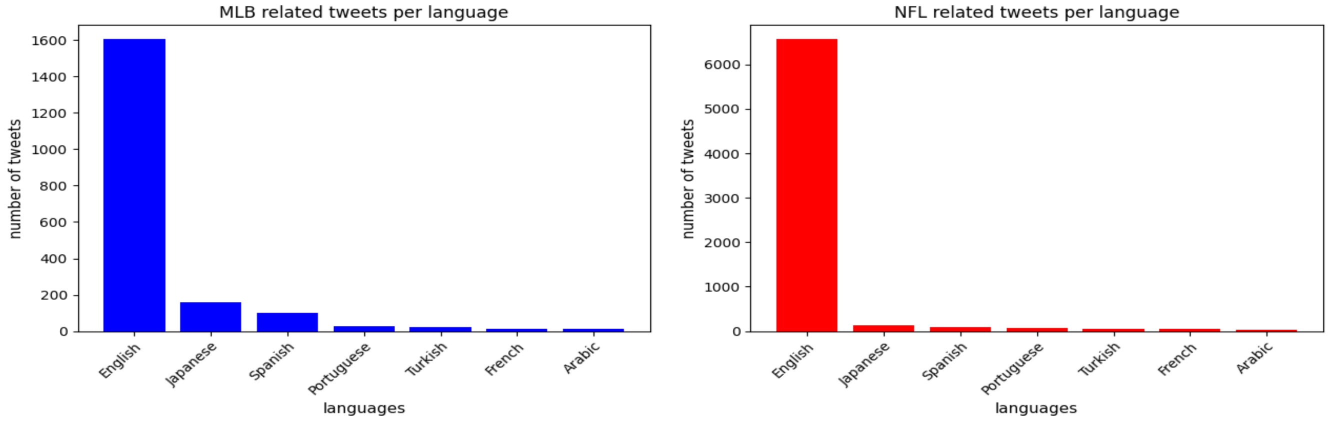
With the data we have and are continuing to collect, we will analyze the volume and engagement sentiment towards sports on social media against the odds and outcomes reported on sports betting sites such as FanDuel and DraftKings.



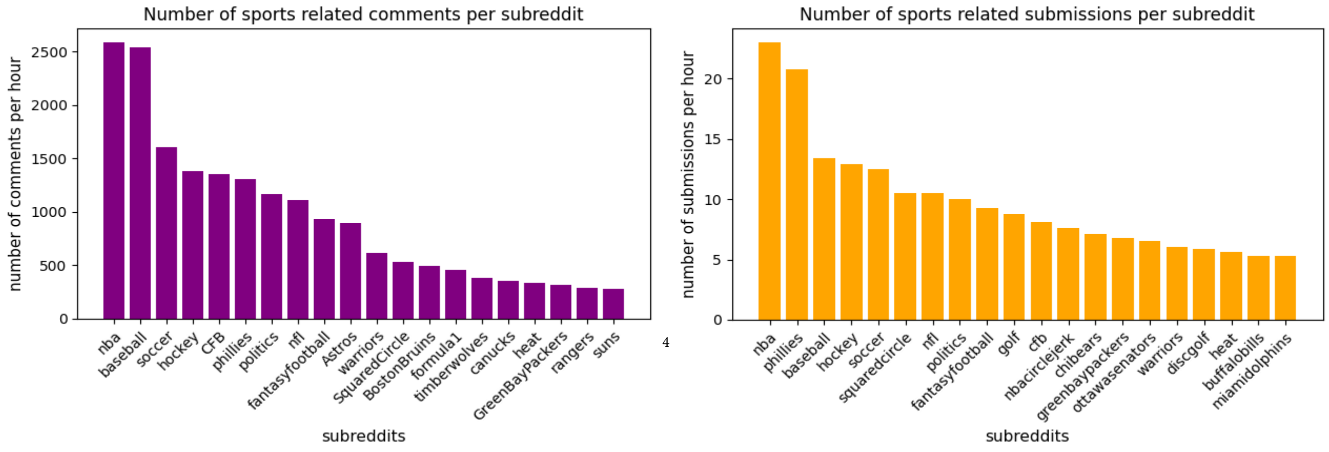
(a) The number of total, stored, and percentage stored tweets per hour



(b) The number of total and stored tweets per language



(c) The number of MLB related and NFL related tweets compared per language



(d) The number of sports related submissions and comments per subreddit