CSCI 49362 Final Research Paper
Can Twitter Help You Win At Fantasy Football?

Lakshmi Palchuri
Hunter College
Lakshmi.Palchuri244@myhunter.cuny.edu

Nuzhat Khan
Hunter College
Nuzhat.Khan86@hunter.cuny.edu

Adam Mohammad
Hunter College
Adam.Mohammad40@myhunter.cuny.edu

Abstract

For our project, we performed sentiment analysis on an unexplored domain of Twitter, fantasy football. While previous sentiment analysis on Twitter data has explored topics like business and politics, we thought it would be interesting to look into a topic about a popular hobby. Specifically, we wanted to determine whether we could rank football athletes according to the sentiment expressed towards them on Twitter and whether this ranking could be used to predict their performance. Using the VADER library, we took a rule and lexical-based approach to sentiment analysis. We found that we could moderately predict rankings, as the average accuracy percentage over the time period observed was 44.17%. However, we believe that there are ways to improve these results, primarily through the analysis of a longer time period.

1 Introduction

One of the most popular social media sites around is Twitter. It is a place that allows users to express their opinions on various topics. These texts, called Tweets, are limited to 280 characters including spaces. For this reason, Tweets tend to be short and concise. Because of the wealth of information contained in Tweets as well as the ease with which words within the Tweets can be tracked, they have been useful for natural language programming research, particularly in sentiment analysis.
Due to the sheer amount of data available, there are still many domains left unexplored. One such domain is fantasy sports, an online activity in which competitors assemble virtual teams of real athletes of a specific sport. These teams compete based on the statistical performance of the athletes in actual games. Fantasy football is the most popular sport, with 39 million of all 59 million fantasy athletes in the US and Canada participating in a fantasy football league (Freedman, 2021). athletes select amongst NFL athletes in a draft in order to assemble their roster. Prior to the draft, they often research to decide which athletes would be best for their team. Similarly, every Thursday, Sunday, and Monday during the NFL season, they decide which athletes should play and which should be benched so that they can maximize their total points for that week. As for the group of athletes who are not drafted at all, called the “waiver wire,” athletes can replace them with one of their drafted athletes at any point in the season.

We wanted to simplify the rigorous process of selecting athletes for teams. Similar to how previous research attempted to predict stock prices from online sentiment, we wanted to see whether Twitter data could be used to evaluate athletes and predict future performance in games (Pagolu, 2017, p. 3). To determine public opinion on specific athletes, we performed sentiment analysis on Tweets about fantasy football. We hypothesized that the positivity or negativity of a Tweet would allow us to rank the athletes, and this ranking could be evaluated by real-time performance rankings. We obtained our data through the Twitter API, gathering 50 Tweets for the top 20 athletes of each week. The tool we used for sentiment analysis is Python’s VADER

library, which provides methods to detect and classify sentiment. In our evaluation step, we ranked the athletes according to sentiment scores and compared this generated ranking to the actual ranking for that week.

2 Previous Research

2.1 Machine Learning Approach

Significant research has been done to explore the applications of sentiment analysis on social media data, mostly using machine learning techniques. One of the hallmark studies on sentiment analysis was done by Pang et al., in which they attempted to reach a high accuracy rate in sentiment classification as in topic classification, which is a comparatively easier task (79). The researchers applied three different machine learning techniques to IMDb movie reviews, which they had categorized as positive or negative according to star or numerical ratings. The machine learning techniques that they used were Naive Bayes classification, maximum entropy classification, and support vector machines (81-82). In order to evaluate their results, they compared the accuracies gathered through these techniques to the baselines three accuracies that they calculated as a preliminary step to the study. They found that the machine learning techniques all exceeded the human, random-choice, and statistical baseline accuracies, but failed to reach the same accuracy seen in topic classification, which could be as high as 90% (84). To explain this accuracy gap, the researchers took a closer look at their dataset and found that there were many "twisted narrative" reviews. These reviews started by listing why a movie should be good or bad but ended by saying it wasn't (85). For future studies, they recommended developing more robust methods that would further incorporate the "understanding" component needed for sentiment analysis.

Our dataset was quite different from the dataset of movie reviews used in this study. Whereas movie reviews tend to be lengthy and verbose, Tweets are short and straightforward. We also were not looking into topics that warranted lengthier Tweets or "Tweet threads." These topics include politics, entertainment, or other news. For football fans on Twitter, a few words generally suffice to express themselves. Adding on to this difference in datasets, training a model would require prelabeled data, but space limitations prevented us from storing large amounts of data. Due to these differences, we looked into rule and lexical-based approaches. A popular tool for this sort of approach is the VADER library.

2.2 VADER Rule & Lexical Approach

According to Gilbert and Hutto, VADER's creators, VADER is perfect for microblogging contexts because it is based on a stellar sentiment lexicon that takes both intensity and polarity into account (Hutto & Gilbert, 2014, p.216). They gathered intensity ratings for their lexicon using ten prescreened human raters who assigned values ranging from -4 to 4 that could be categorized as absolutely negative [-4], neutral [0], or positive [4] (p. 220). To increase the sensitivity to sentiment, they implemented five heuristics. These were punctuation, mainly exclamation points; capitalization, specifically fully capitalized words among other lowercase words; degree modifiers, such as 'extremely' and 'marginally'; 'but' for shifts in sentence polarity; and negations caught using trigrams. Through a combination of quantitative and qualitative techniques, the creators determined values for the effects that these heuristics would have as well as the scope of their effect (221). After evaluating VADER, they found that the heuristics vastly improved the accuracy of sentiment analysis in several contexts (e.g.

movie reviews, news articles, social media, etc.) tested, especially for social media data (222). VADER performed well against human raters and even outperformed individual raters. Its performance was also compared to eleven other benchmarks, including the same machine learning techniques used in Pang et al.'s study. It held its own when compared to these other more sophisticated approaches (224).

2.3 VADER Application

Elbagir and Yang's study attests to the effectiveness of VADER. They did sentiment analysis on Tweets about the 2016 presidential election (12). Most Tweets were either negative or neutral, and the sentiment analysis reflected this to a degree since the number of Tweets identified as neutral was almost double that of Tweets identified as negative and positive (14). Interestingly, despite the small number of positive Tweets in the pre-labeled data, the percentage of Tweets identified as positive was nearly equivalent to the percentage identified as negative. The researchers note that a limitation of their study was the small number of Tweets used (15).

We take a similar direction as this study by using VADER for sentiment analysis of fantasy football Tweets.

3 Data

3.1 Weekly Athletes' List

Fantasydata.com is a reputed site among fantasy football athletes. It has a detailed database of athletes in the NFL. From this site, we manually selected a list of the top twenty athletes every week. Their names and their current rankings were noted down so we could use this information in our evaluation.

3.2 Athletes' Tweets

We scraped the Tweets mentioning the handpicked athletes from Twitter through open API calls. We used the Twitter API along with the Tweepy library to fetch Tweets that mentioned the names of athletes that we previously identified in our weekly athletes' list. Essentially, every week we would identify a list of athletes from the top of FantasyData.com's ranking database. Our script iterated through this list and used Tweepy to obtain Tweets mentioning each of the athletes. Every Tweet we collected was a string. Using a dictionary, we grouped each athlete's Tweets with their name so we could easily identify which Tweets belonged to which player. We encountered a challenge once we realized that our Twitter API access was limited to only access Tweets posted over the last week. We decided to collect 50 Tweets for each of the 20 athletes each time we ran the script. Since we ran the script three times a week, we collected 3,000 Tweets each week. In total, we processed 12,000 Tweets. After we collected each week's Tweets, we preprocessed and cleaned them.

4 Method

The input for this experiment was a dataset of Tweets separated by athlete name and the output we wanted was two lists of athletes classified as "strong" athletes and "weak" athletes. We used a three-step implementation for our approach (adapted from D'Andrea et al., 2015, p. 26):

1. The first step was sentiment detection, which we did use VADER's rule and lexicon-based sentiment analysis. Each Tweet was given a sentiment score which included how positive, negative, and neutral a sentence was along with an aggregated compounded score.
2. The second step was sentiment classification where we used the compounded score calculated previously and found how many positive/negative/neutral Tweets each athlete had on average.

3. The last step was the presentation of output. We took the data produced by the previous step and performed different forms of evaluations on them to be able to group the athletes in the two classification categories.

4.1 Preprocessing

Before we started the steps that were listed out above, the Tweets we collected needed to be preprocessed. All mentions (ex. @name) and URLs were removed, all digits were removed, all emojis were converted to text, and all non-negative stop-words were removed. To note, we used NLTK's stop-words list as the basis for our stop-words list but modified it to remove certain ones that we believed were necessary for sentiment analysis. Primarily, we wanted negated words (e.g. don't, won't) to remain in our dataset. The stopwords that were removed were words we believed would be uninformative words. We also compiled a list of the ten most and least frequent words from all the Tweets. Then we removed these words if they existed in a Tweet. We did so because we wanted this step to remove uninformative and rare domain-specific words. Lastly, we applied lemmatization by applying NLTK's wordNetLemmatizer. We refrained from applying some standard preprocessing practices like removing punctuation and making all the text lowercase because our chosen sentiment analysis tool used punctuation and capitalization heuristics to determine sentiment.

4.2 Sentiment Analysis

After preprocessing, we applied to each athlete's cleaned Tweets. VADER assigns a word score to every word in the Tweet then sums and normalizes these word scores. This produces an overall sentiment score for each Tweet which can range from -1(negative) to 1(positive). As mentioned in the 'previous research' section, the punctuation heuristic amplifies the sentiment score for single or repeated punctuation (e.g. '!!!!'). The capitalization heuristic similarly dealt with sentiment. For degree modifiers, the scores of the words following the modifier are decreased or increased. This effect decreases gradually for each word following the modifier. Lastly, for the negations we decided to keep, if VADER identifies a negation through its use of trigrams, the sentiment score is amplified.

Following sentiment analysis, we classify the Tweets as positive, negative, or neutral using the compounded score VADER calculated. We kept a running total for the number of positive Tweets each athlete had and we did the same for negative and neutral Tweets. After calculating the number of Tweets in the three sentiment categories, we took the average to determine the mean percentage of positive, negative, and neutral Tweets each athlete had. While doing this, we noticed that every time we ran the script, the Tweets that we were gathering were different. Since we were unable to save the data from each round of analysis, we decided to counteract this potential source of error by running the script three times per week (every Monday, Tuesday, and Wednesday to get a more widespread sample of Tweets for each athlete) and manually taking the average of the results we gathered for each athlete. We also calculated athlete-specific positive, negative, and neutral Tweet counts to see if we can observe any trends. We performed evaluation methods on these results. To visualize the results, we created scripts to generate graphs (fig. 1-8). We also kept track of each week's data through a comprehensive set of tables on Excel.

5 Results/Evaluation

To determine whether sentiment is a worthwhile metric in evaluating fantasy

performance, we sought to find a correlation between Twitter sentiment and the top-performing fantasy athletes week-by-week. Our experiment spanned four weeks beginning on November 22nd, 2021, and ending on December 13th, 2021. At the conclusion of each Monday Night Football game, which marks the end of the fantasy week, we compiled a list of the top-20 fantasy performers. That is the athletes that scored the most fantasy points each week. Every week, we ran our script three times, gathering thousands of Tweets that mentioned each athlete's name and were Tweeted prior to the Thursday night football game, which marks the start of the fantasy week. This represented the latest public sentiment for athletes right before football games began. As mentioned before, our script then outputted the percentage of positive, negative, and neutral Tweets pertaining to each athlete. This data was placed in a spreadsheet each week. We then manually put each player's average polarity percentages into another script which outputted additional lists sorted according to:

- Highest percentage of positive Tweets
- Lowest percentage of negative Tweets
- Highest percentage of neutral Tweets
- Highest difference between positive and negative Tweet percentages
- Highest ratio of positive to negative Tweets
- Highest percentages of positive + neutral Tweets

We explored multiple methods of sorting to see how the rankings differed and whether or not certain methods would consistently lead to more accurate rankings. We observed how often the sentiment ranking, according to each of the above metrics, grouped athletes into the correct half of the top-20 fantasy performers list.

In the first week, which was week 11 of the fantasy season, the average accuracy of correct groupings using all metrics was **38.33%**. Next week, week 12, the total average accuracy was **46.67%**. In the following week's group of athletes, we received a similar accuracy of **45%**. In the final week, week 14, we still calculated a similar accuracy of **46.67%** (fig. 9). We also ran the same script on the next 20 athletes, for comparison's sake, and our script grouped athletes ranked 21-40 with an accuracy of **54%**.

Overall, across all four weeks, our script had a total average accuracy of **44.17%**. In other words, using sentiment analysis, we grouped the list of each week's top-20 fantasy athletes into the correct half of the list 44.17% of the time.

For a more precise gauge of correlation, we also calculated the Kendall-tau rank coefficient, a measure of the similarity between two ranked lists, our sentiment rankings, and the actual real-world rankings. The coefficient values are numbers ranging between -1 (indicating complete disagreement between the lists) and 1 (indicating perfect agreement). Values near zero indicate no correlation between the two lists. Our calculated coefficients were **0.008333, 0.132353, 0.015790,** and **0.001755** for Weeks 11 through 14, respectively. The total average correlation across all four weeks was **0.039558,** which expresses low correlation between our sentiment score rankings and the actual real-world rankings.

Consequently, our experiment demonstrated that there is a weak relationship, if any, between opinions expressed on Twitter about fantasy football athletes and the real-world performance of those athletes. However, this project is limited in scope and subject to several sources of error. The most prevalent of which is the temporal nature of data

collection. According to Twitter, 500 million Tweets are sent every single day and almost 6,000 are sent every second, meaning that the timing of collecting Tweets may impact results ("*New Tweets per second*", 2013). As we noted before in our experiment, every time our script was run, a new corpus of Tweets was generated with different sentiment data. Though we ran the script multiple times a week and averaged the sentiment percentages, this is still an imperfect solution.

Another source of error involves the process of searching for Tweets that correspond to each athlete. While our search method scraped Tweets that mentioned an athlete's full name, it's more common for users to refer to professional athletes by the last name or with nicknames. Initially, we searched for an athlete's last name and full name within Tweets; however, this cluttered our corpus with Tweets that referred to any person that shared a last name with the NFL athlete. Last names that were English words were another complication using this method. For example, searching for Dalvin Cook and Ja'Marr Chase would return Tweets that mentioned the words "Cook" and "Chase." As for the nicknames, many athletes such as Ben Roethlisberger, Adrian Peterson, and Matt Ryan are more commonly called: "Big Ben," "AP", and "Matty Ice," respectively. Nicknames for athletes are especially prevalent on informal platforms like Twitter. To ensure the relevancy of Tweets, we gathered Tweets that mentioned the full names of athletes, but this neglects a large portion of Tweets.

We also noticed that some Tweets that were gathered were simply advertisements that just stated the athlete's name and a link to a website. More often than not, these kinds of Tweets were rated as neutral. They cluttered our dataset and inflated the number of neutral Tweets directed towards a player.

Neutral Tweets were another significant challenge of this study. Most of the previous research that we looked into discounted neutral texts. Some mentioned that although neutral Tweets were discounted, there should be some sort of accountability for them (Go et al., 2009, p. 151). Since VADER calculates neutrality, we did rank according to neutrality as well. We found that the accuracies were a bit higher according to this ranking, but since neutral Tweets don't express meaningful sentiment in the context we explored, we believe this slightly higher accuracy was likely due to a random chance. Of course, analyzing across a greater time period and with much more data would help to confirm this.

Neutrality also comprised the majority of determined sentiment. In this regard, our results were similar to Elbagir and Yang's study, in which they found that most Tweets were identified as neutral (14). Just as they recommended analyzing a much greater dataset, we believe the effect of neutrality on our dataset can be reduced by analyzing a lot more Tweets.

6 Conclusion

Through this project, we were able to apply sentiment analysis on an uncharted domain of Twitter. We used a rule and lexicon-based approach developed by the VADER library to perform the sentiment analysis. By manually ranking the athletes according to the sentiment scores we gathered and comparing this ranking to the true ranking for that week, we were able to evaluate whether or not sentiment scores could be used as a predictor for athlete performance. Over the course of four weeks, the average accuracy percentages steadily grew, though we cannot say with absolute confidence that there was a positive trend due to the short timeframe we observed. The overall average accuracy rate for the four weeks was 44.17%. We believe there are

several ways to improve upon this score, including analyzing across a much longer time period, gathering more data, and perhaps applying machine learning techniques for comparison purposes.

7 Future Work

For the best results, this experiment should be conducted throughout a full-length fantasy season, spanning seventeen weeks. Additionally, search parameters must be individually defined for each athlete, taking into account the most common names or nicknames they are referred to. Lastly, the method of evaluating results should be refined, perhaps by grouping different tiers of athletes and evaluating their performance within those tiers, or evaluating the performance of different groups of athletes against other groups of athletes.

Additionally, further research is needed on the unique world of sports and social media. For example, although our project was focused on Twitter, would sentiment analysis on other social media sites produce similar results? If not, which site is the most accurate? Additionally, examining the way that people communicate about sports online is another area where further research can be interesting. For example, can social media data support the existence of a positive or negative bias towards athletes of a certain race over other races? Overall, it is the authors' belief that the widely unexplored cultural intersection of sports and social media can provide researchers with compelling insights into humans in general.

References

D'Andrea, A., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, *125*(3), 26–33. https://www.ijcaonline.org/research/volume125/number3/dandrea-2015-ijca-905866.pdf

Elbagir, S. E., & Yang, J. Y. (2019). Twitter sentiment analysis using natural language toolkit and VADER sentiment. *Proceedings of the International MultiConference of Engineers and Computer Scientists 2019*, 12–16. http://www.iaeng.org/publication/IMECS2019/IMECS2019_pp12-16.pdf

Go, A., Bhayani, R. & Huang, L. (2009). Twitter sentiment classification using distant supervision. *Processing*, 150-155. https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

Hutto, C. J. H., & Gilbert, E. G. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Mediatext*, 216–225. https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. Z. (2011). Target-dependent Twitter sentiment classification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 151–160. https://aclanthology.org/P11-1016.pdf

Joshi, P. (2021, December 17). RBO v/s Kendall Tau to compare ranked lists of items. *Medium*. https://towardsdatascience.com/rbo-s-Kendall-tau-to-compare-ranked-lists-of-items-8776c5182899

Kim, S. M., & Hovy, E. (2004). Determining

the sentiment of opinions. *Proceedings of the COLING Conference*. Published. https://www.cs.cmu.edu/~hovy/papers/04Coling-opinion-valences.pdf

Mai, L., & Le, B. (2020). Joint sentence and aspect-level sentiment analysis of product comments. *Annals of Operations Research*, *300*(2), 493–513. https://link-springer-com.proxy.wexler.hunter.cuny.edu/article/10.1007/s10479-020-03534-7

New Tweets per second record, and how! (2013, August 16). Engineering. Retrieved December 20, 2021, from https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how

Pagolu, V. S. (2016, October 28). Sentiment analysis of Twitter data for predicting stock market movements. *ArXiv*. Retrieved October 20, 2021, from https://arxiv.org/abs/1610.09225

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of EMNLP*, 79–86. https://www.cs.cornell.edu/home/llee/papers/sentiment.pdf

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. 417
down? Semantic orientation applied to unsupervised classification of reviews. 417-–424. https://aclanthology.org/P02-1053.pdf

Contributions

Data Collection

Adam gathered all needed data from fantasydata.com and gathered the Tweets from Twitter using Tweepy

Text Preparation

Lakshmi preprocessed the text/Tweets

Sentiment Detection

Lakshmi used VADER to perform sentiment detection on each athlete's Tweets

Nuzhat attempted NLTK machine learning sentiment detection but did not get good results.

Sentiment Classification

Lakshmi, Nuzhat, and Adam used the sentiment analysis data to classify athletes meaningfully and produce useful statistics

Presentation of Output

Nuzhat and Adam formatted the outputs yielded in our program into tables and graphs

We all contributed to the experiment scripts and worked on the research paper.

Nuzhat wrote the Abstract, Introduction, Prior Research, and Conclusion sections. Edited other sections.

Lakshmi wrote the Methods and Data Collection sections.

Adam wrote the Results/Evaluation section along with the Future Works section

Appendix

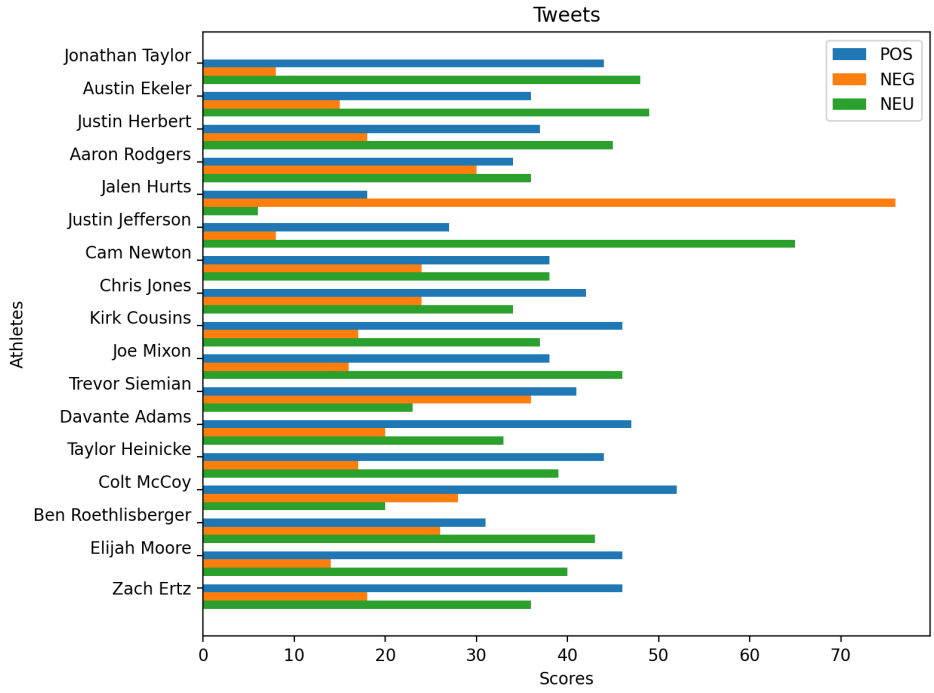

Fig 1. *Avg sentiment scores for top athletes week 11*

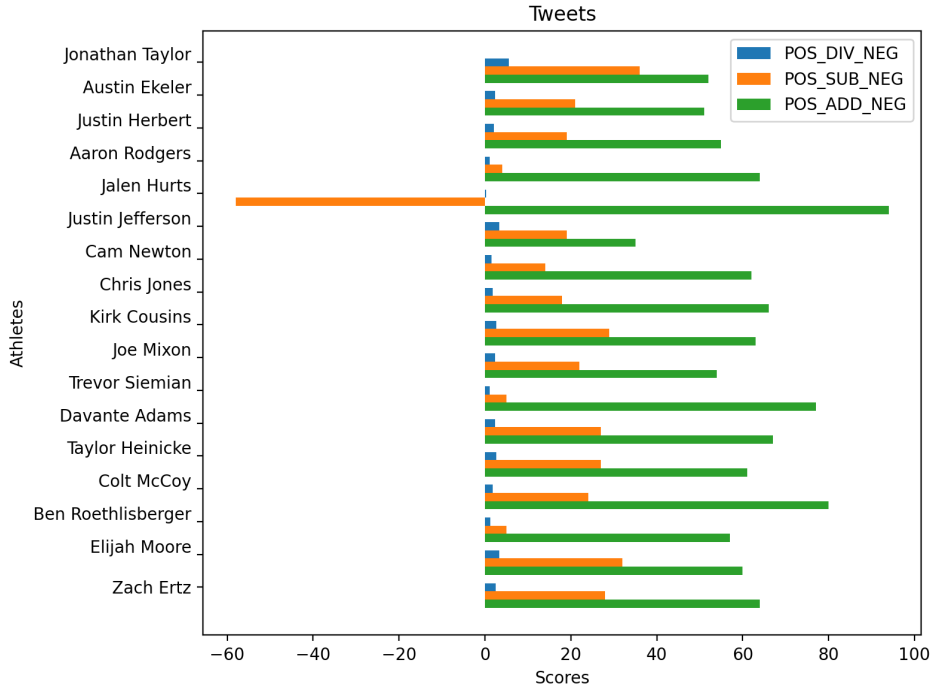

Fig 2. *Avg sentiment scores for top athletes using different metrics*

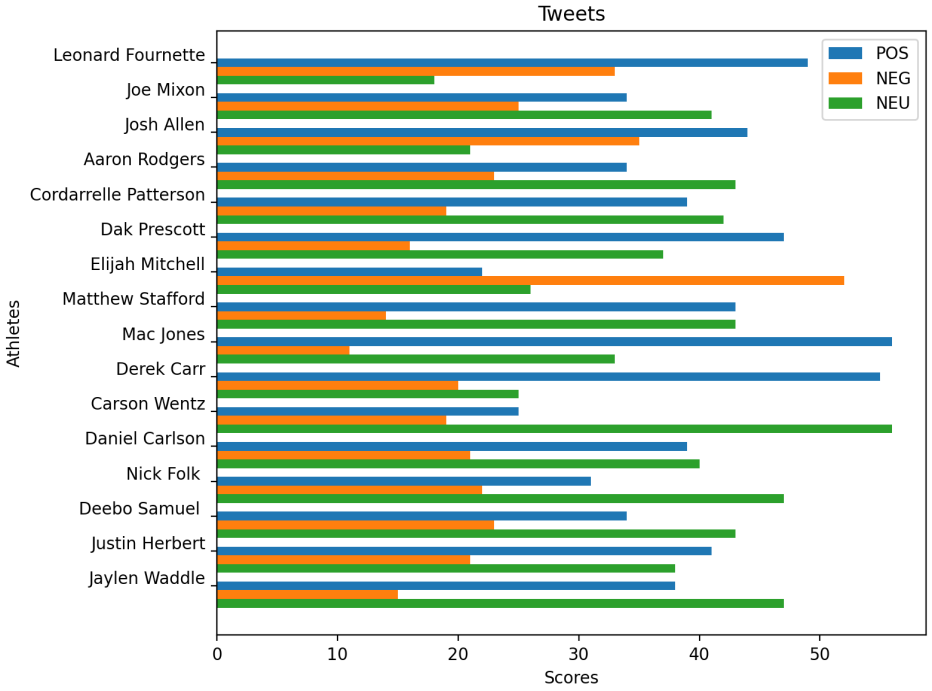

Fig 3. *Avg sentiment scores for top athletes*

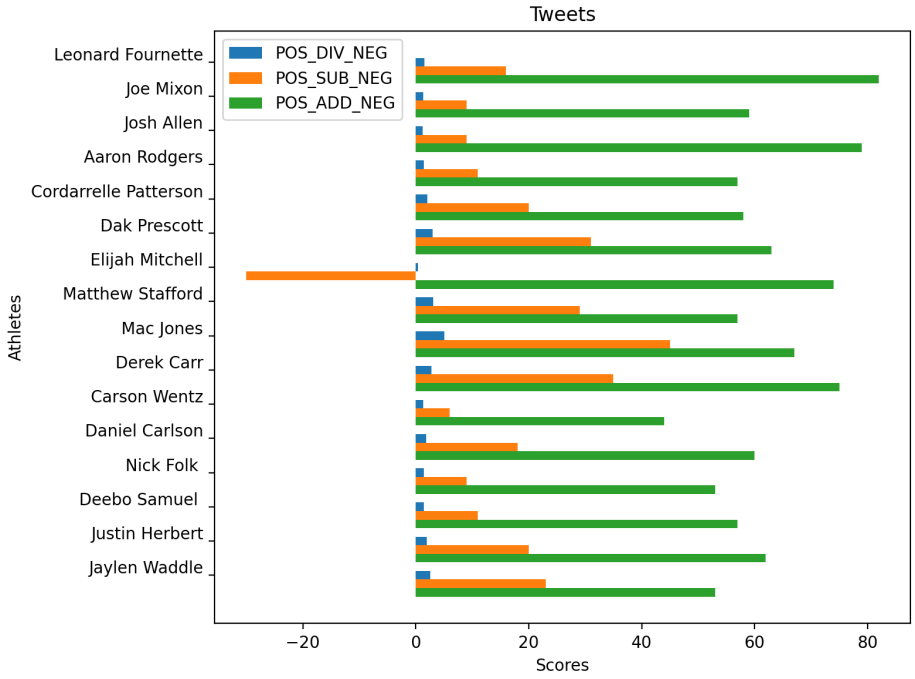

Fig 4. *Avg sentiment scores for top athletes using different metrics week 12*

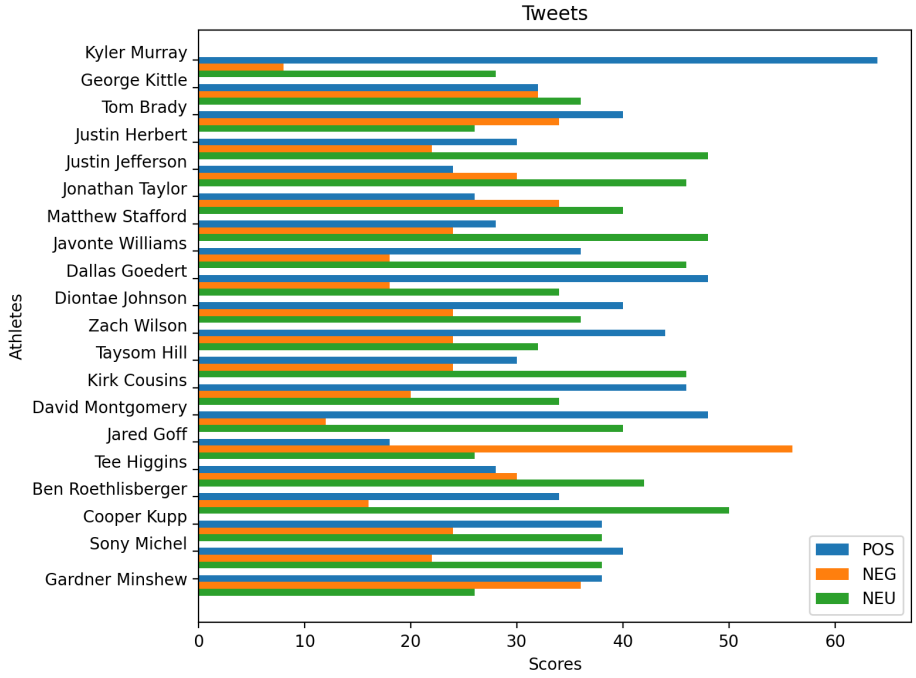

Fig 5. *Avg sentiment scores for top athletes week 13*

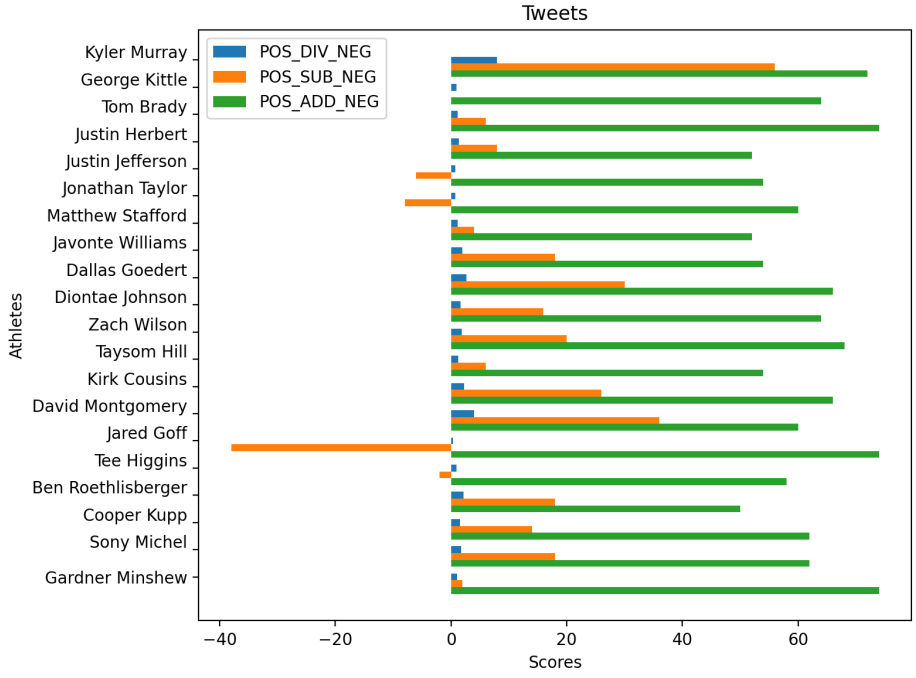

Fig 6. *Avg sentiment scores for top athletes using different metrics week 13*

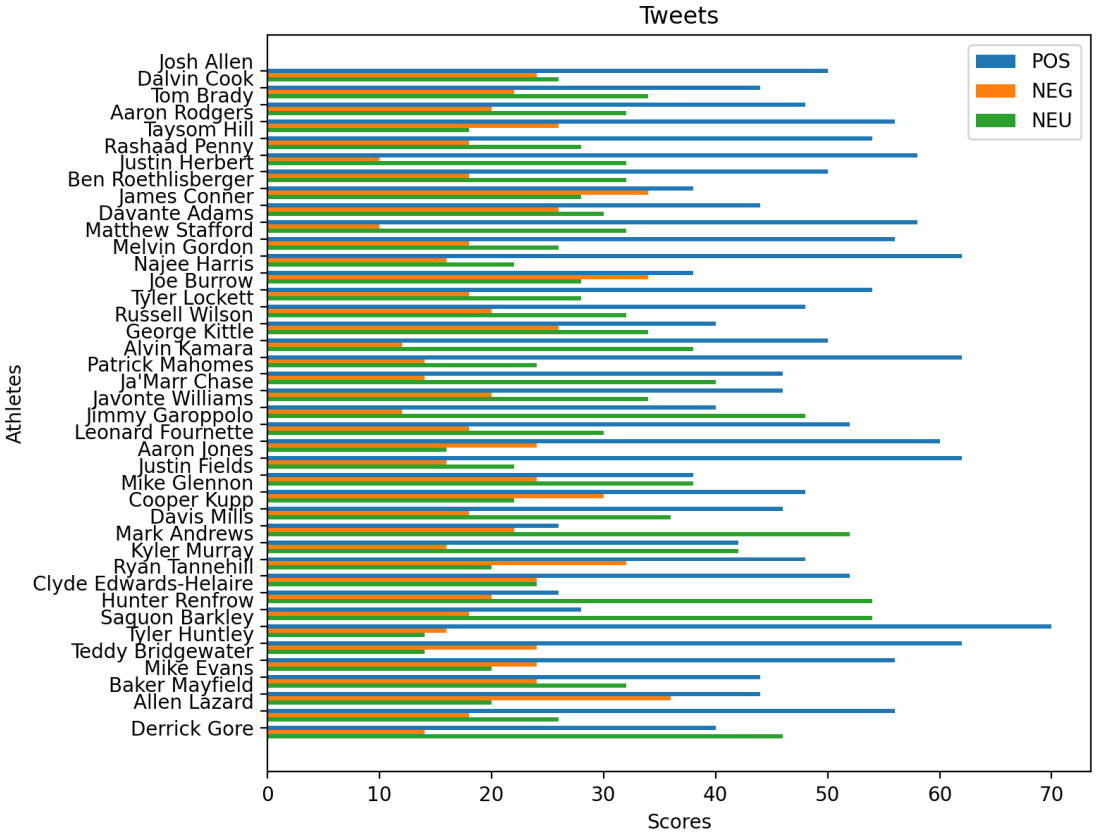

Fig 7. *Avg sentiment scores for top athletes week 14*

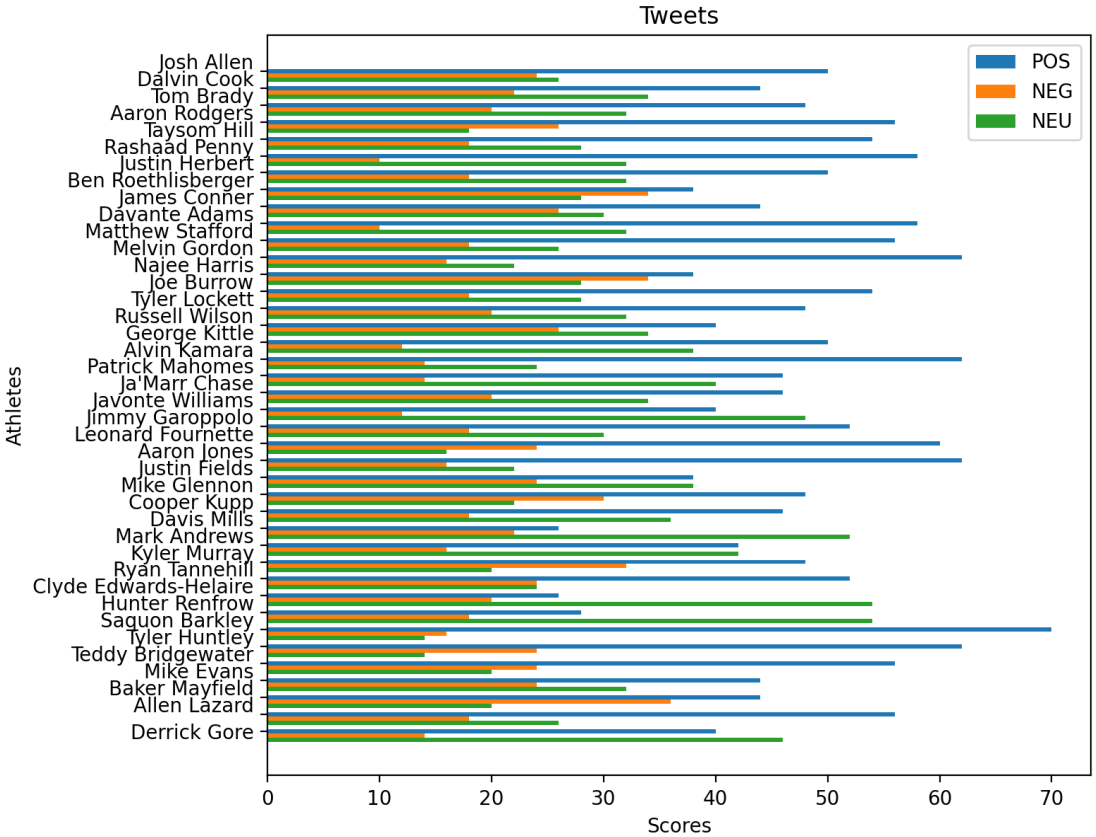

Fig 8. *Avg sentiment scores for top athletes using different metrics week 14*

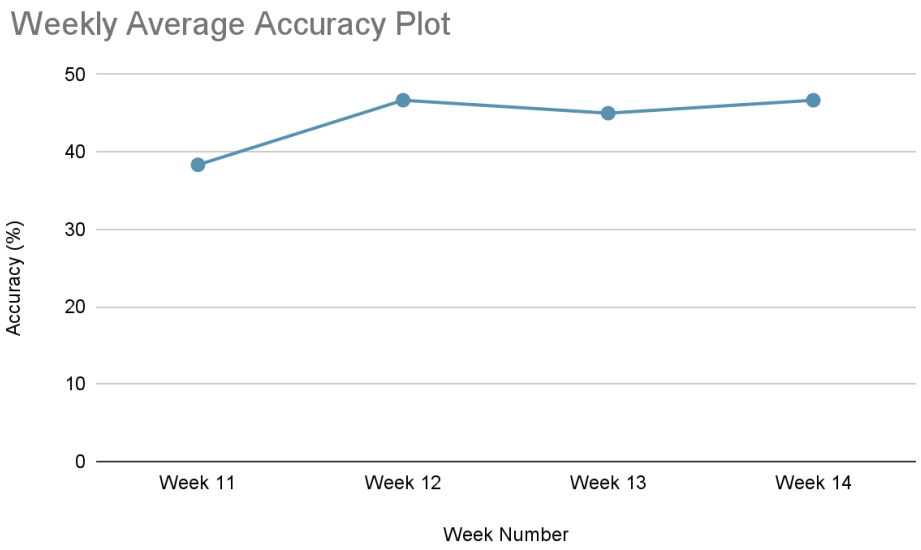

Fig 9. *Avg accuracy percentages over 4 weeks*