Sentiment Analysis of the NBA SubReddit

Ricardo Cruz
Portland State University
Portland, Oregon
rcruz@pdx.edu

Abstract - Sentiment analysis can inform decision making and is often used in domains such as social media. The NBA subreddit is an online community where users discuss NBA teams and players. The top posts from Post Game Threads in the NBA subreddit were selected for analysis using Valence Aware Dictionary for sEntiment Reasoning (VADER) and TextBlob generated sentiment analysis. A comparison on the generated sentiment polarity scores between the two tools showed VADER scored posts more negatively than TextBlob. The most posts with a negative sentiment belonged to the Los Angeles Lakers and positive sentiment to the Chicago Bulls. The Cleveland Cavaliers held the highest ratio of positive posts, and player specific sentiment analysis showed a member of their roster, Even Mobley, had the most posts with a positive sentiment. Russell Westbrook held the most posts with a negative sentiment. These findings hint that VADER is better suited for this domain. Sentiment analysis in this domain can be combined with other metrics to make predictions on outcomes between two teams along with informing interested parties which players and teams are driving certain engagements.

1 Introduction

This paper proposes a data-driven approach to comparing the sentiment polarity scores generated by Valence Aware Dictionary for sEntiment Reasoning (VADER) and Text Blob in an online community.

Reddit is a web platform and can best be described as a platform for social news aggregation and discussion. There are a variety of topical discussion forums, or subreddits, allowing users to submit text posts or direct links to be curated by others. Reddit has been used in studies by researchers for Natural Language Processing because of how it is topically constructed. Such research includes those in the field of categorising biases and sentiment analysis [1].

There are a variety of held interests in major sports such as advertising towards key demographics, fantasy sports, sports betting, and media coverage to name a few. Determining the sentiment held for a certain team can assist individuals whom participate in sports betting gain an advantage if a team has a positive sentiment while a negative sentiment could suggest a team is under-performing and should be avoided in such a scenario. Social media has gained interest in its use to make predictions over a diverse number of domains because of its crowd-based content that can collected and used as a dataset for analysis[7].

Over 4.2 million members are subscribed to the NBA subreddit with traffic averaging 20 thousand users or more online The sentiment polarity scores for comments in the NBA subreddit generated by both VADER and TextBlob will be compared to establish which is better suited for the domain.

2 Background

2.1 Sentiment Analysis

Sentiment analysis can be applied to a wide range of domains such as sociology, marketing and advertising, and economics. Public engagement and sentiment towards brands, products, or events present opportunities for opinion extraction. The resulting insights into public sentiment can be used to proactively to inform decision making [5].

Many of the approaches rely on the underlying lexicon used to generate sentiment polarity scores. Challenges to practical applications of sentiment analysis exist with social media content. The tendency to use abbreviated language to express sentiment creates contextual sparseness.

This is often observed on social media platforms such as Twitter and Facebook. Large, high quality lexicons are important for accurate sentiment analysis, but the researchers argue existing lexicons are often used in less suitable domains like social media [3].

2.2 VADER

Hutto and Gilbert are responsible for the development, validation, and evaluation of a simple rule-based model for general sentiment analysis called VADER (Valence Aware Dictionary for Sentiment Reasoning) to address this existing problem of insufficient lexicon use in the social media domain. The intent behind this research was to construct a computational sentiment analysis engine that readily works well on social media style text, requires no training data and is instead constructed from a valence-based sentiment lexicon, does not suffer from a speed-performance trade-off, and is fast enough to use with streaming data online.

VADER was also found to have advantages over machine learning techniques that make it faster and computationally economic. A large corpus can take a substantial amount of time to analyze using SVM while VADER is able to process it at fraction of the time. This contribution does not sacrifice accuracy at the expense of speed, making VADER's rule-based model an efficient alternative for usage in social media domains [4].

2.3 TextBlob

TextBlob's Sentiment Classifier is a modular approach and by default uses a Naive Bayes classifier trained with unigram features. TextBlob is a library included in Python's Natural Language Processing library and tool kit that has historically been used to generate sentiment polarity scores for sentiment analysis of content on social media platforms like Twitter [2]. One such use case was a sentiment analysis of the Twitter handle of President Donald Trump where tweets were compared before and after the 2016 United States Presidential Election. TextBlob was found to be highly efficient for use with large datasets [6].

3 Methodology

3.1 Data Extraction

PRAW, the Python Reddit API Wrapper was used to build the dataset consisting of posts submitted to the NBA subreddit. This required configuring a custom request and the registration of a Reddit account.

There are *Game Threads* regularly created at the start of each NBA game, but these often contain short, reactionary comments that are not insightful. Instead, *Post Game Threads* were selected to use to build the dataset. While these threads often contain reactionary comments as well, they are generally more thoughtful and analytical in nature.

The 2021-2022 NBA season began October 19, 2021, and the date served as the starting point of the dataset. November 17, 2021 was the final date, thus providing 30 days worth of *Post Game Thread* data.

Next, the NBA schedule for the selected range of data was obtained from www.basketball-reference.com. Additionally, all 30 team names and up-to-date rosters with the names of the players for each of the respective teams were stored.

With the schedule data now obtained, 222 corresponding game threads were accessed in order to store the unique ID used by PRAW to obtain the top posts. Along with the ID, the respective participant teams and the over 15,000 posts obtained using PRAW were stored to serve as the primary dataset.

3.2 Data Processing

The dataset was cleaned by first dropping rows where posts were deleted. These posts are not omitted by Reddit during the retrieval process. They must be removed during data cleaning because they will register a sentiment polarity score otherwise.

Next, the case for the text was adjusted to be in lower case. Hashtags, links, and special characters were also removed from the text. This type of data in the posts will cause unexpected results from the sentiment analyzer and could affect the accuracy of the sentiment polarity scores due to irrelevant context. Multiple blank spaces were then replaced by single spaces after the text was cleaned.

4 Experimental Setup

4.1 Lexicons

Python's NLTK and TextBlob library were used in the experiment. NLTK's VADER lexicon coupled with VADER's Sentiment Analyzer was used to generate sentiment polarity scores. The text was

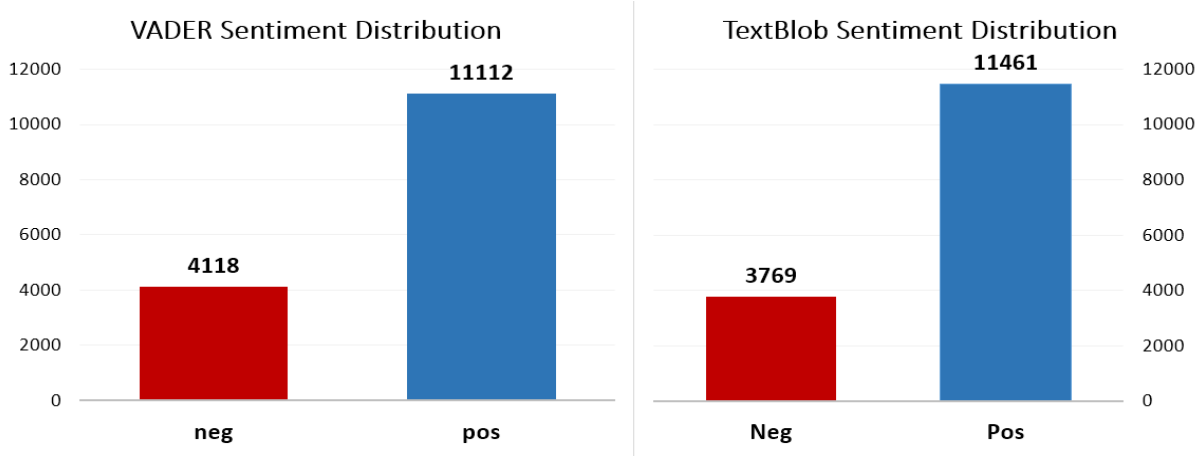

Figure 1: Distribution of polarity over all posts

independently subjected to a TextBlob instance and the Sentiment Analyzer, where a sentiment polarity score was obtained from the respective tool.

For the VADER analysis, a negative, neutral, and positive score was assigned to each post. A compound score was then calculated by taking the sum of each post. Standardized thresholds were used when determining how to classify both respective scores. When the score ≥ 0.05 the sentiment was classified as positive. When the score ≤ -0.05 the sentiment was classified as negative. A neutral sentiment classification is defined as score < 0.5 and score > -0.5 and was omitted from the experiment.

4.2 Targets

A sentiment polarity score was generated for every post in the dataset using both VADER and TextBlob, respectively. Since each post was tagged with the respective participant teams, simply filtering for each of the 30 teams would yield a sentiment analysis for the *Post Game Threads* they were involved in. A more specialized approach towards filtering was implemented.

Each team's respective roster was iterated by player name where the name served as the target. Any post containing the target was stored with the target, corresponding positive or negative sentiment label, and the team name. The team name also served as the target for each of the 30 teams. When the filtering was complete, the dataset was reduced to only posts mentioning any of the players by name or the team by name.

Targeting specific posts thus enabled for team specific sentiment analysis along with player specific sentiment analysis in *Post Game Threads* from the NBA subreddit.

5 Results

5.1 Distribution

After performing sentiment analysis over all the posts in the dataset, VADER vs TextBlob distribution (Figure 1) revealed VADER's threshold for scoring posts positively was higher than that of TextBlob. Further review of some of the types of posts where this discrepancy is apparent can be found in Table 1.

TextBlob	Vader	Post
Pos	Neg	dont want to overreact but i think the nets arent making the playoffs
Pos	Neg	bruh we suck
Pos	Neg	tbf to westbrook while he had 7 tos if we look at the ones that were actually his fault then he really only had 11 tos

Table 1: Post Sentiment

TextBlob showed more difficulty scoring sarcasm and irony in this domain than VADER. The first row of the table contains a post about a team missing the playoffs. This type of post would likely be classified as negative by most basketball fans, but TextBlob scored it as a positive sentiment.

5.2 Teams

Sentiment analysis with both VADER and TextBlob revealed the Los Angeles Lakers to be the team with the most posts with negative sentiment. Rounding out the top 4 were the Chicago

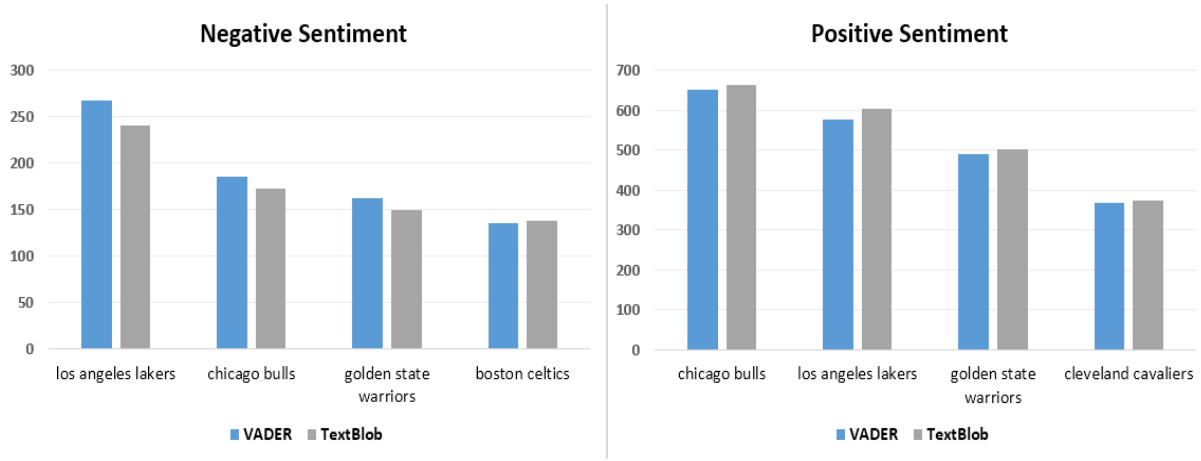

Figure 2: Sentiment Analysis of Top 4 Teams

Bulls, Golden State Warriors, and Boston Celtics (Figure 2). VADER scored more team specific posts with negative sentiment than TextBlob did, except for the Boston Celtics.

The top 4 teams with the most posts with a positive sentiment were the Chicago Bulls, Los Angeles Lakers, Golden State Warriors, and Cleveland Cavaliers. TextBlob scored posts more positively than VADER did.

The teams with the most posts scored were the Los Angeles Lakers, Chicago Bulls, Golden State Warriors, and the Cleveland Cavaliers (Figure 3). VADER also scored these team specific posts more negatively than TextBlob.

The results for team specific sentiment analysis was consistent with the initial distribution over all the posts that found VADER to score more negatively than TextBlob.

5.3 Players

Sentiment analysis with both VADER and TextBlob revealed Russell Westbrook to be the player with the most negative sentiment. Additionally, Nikola Jokic, James Harden, and Luka Doncic were in the group of top 4 players with the most posts with negative sentiment (Figure 3). VADER scored more posts negatively than TextBlob except for Luka Doncic where TextBlob outscored negative posts than VADER.

The top 4 players with the most posts with a positive sentiment were Evan Mobley, DeMar Derozan, Stephen Curry, and James Harden. TextBlob scored posts more positively than VADER in this instance except for Stephen Curry where VADER scored posts more positively.

Generally, the results for player specific sentiment analysis was also consistent with the findings of team specific sentiment analysis. They both followed the initial distribution over all the posts where VADER was shown to score more negatively when compared to TextBlob.

6 Conclusion

6.1 Limitations

This experiment was not without limitations. Targeting player names omits posts where the name is shortened. Popular players like Stephen Curry or Carmelo Anthony are rarely referenced by their full first name. Often times the name is shortened to Steph or Melo. Different players also go by nicknames. Pascal Siakam is referred to as Spicy P and Giannis Antetokounmpo is often called Greek Freak. These names were not targeted and the posts did not make it past the filtering process. Another limitation during the targeting process was abbreviations of players or team names. Kevin Durant is often referenced as KD. A team like the Golden State Warriors is abbreviated to GSW.

There were also many posts sharing a general sentiment about the game or a reaction to the context outside of player or team performance. Posts about the crowd, referees, fouls, and coaches were present in the dataset. General sentiments about how the game made the user feel was not relevant to the targeted sentiment analysis, so many of these posts were omitted.

Using pre-built lexicons were not best suited for the project. It is suspected VADER performed

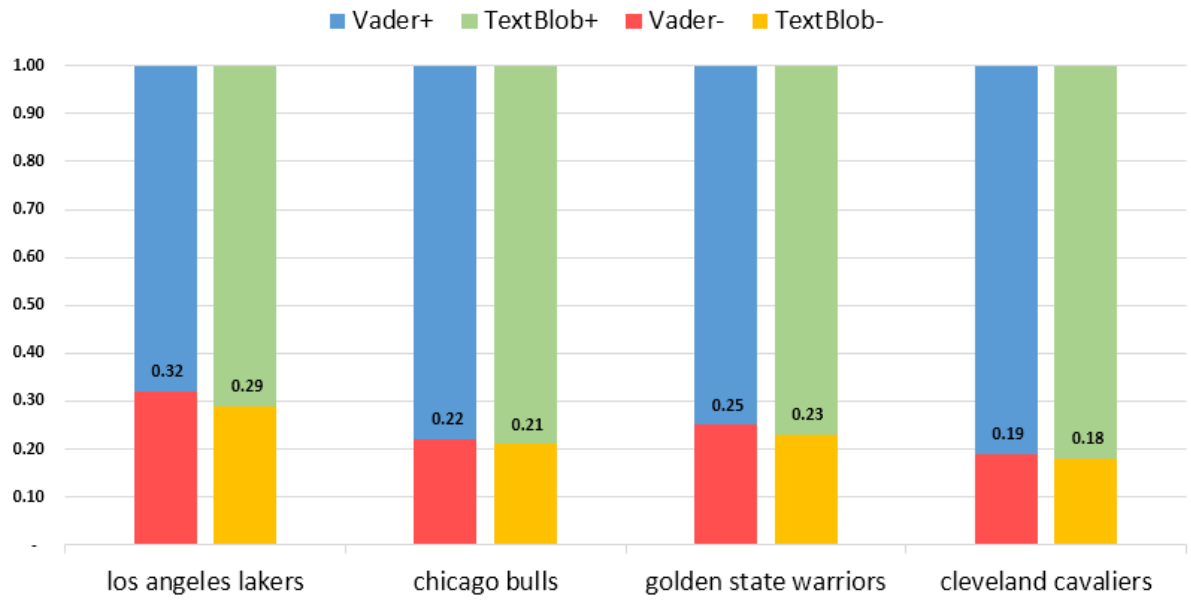

Figure 3: VADER vs TextBlob Sentiment Analysis of the Top 4 Teams With the Most Posts

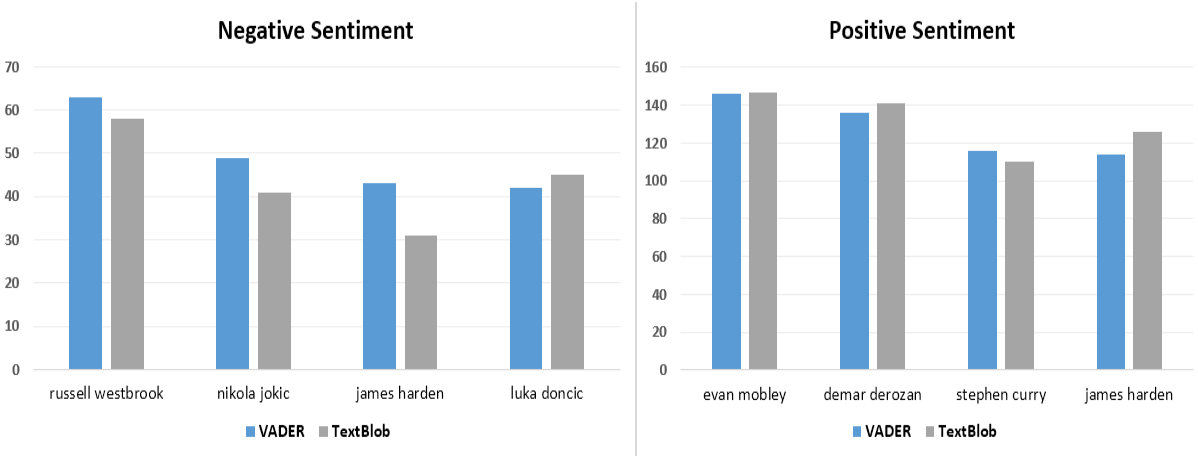

Figure 4: Sentiment Analysis of Top 4 Players

better since it was created to be used in the social media domain. However, there were instances where basketball related context was missed by both VADER and TextBlob.

6.2 Notable Findings

Results showed that VADER was better suited for sentiment analysis of the NBA subreddit. TextBlob scored posts more positively and displayed a tendency to miss irony and sarcasm. That's not to say VADER did much better, but VADER was able to catch critical posts and label them correctly at a better rate than TextBlob.

The results for the team sentiment analysis were surprising. Negative sentiment was highest for the Los Angeles Lakers and it was suspected since they are a large basketball market. Interestingly, Russell Westbrook had the most posts with a negative sentiment than any other player. He also happens to be a member of the Los Angeles Lakers, so he contributed significantly to the negative sentiment of the team as a whole. Another interesting insight was how many posts with a positive sentiment were present for the Cleveland Cavaliers. This instance was analogous to the previous one of the Lakers because the player with the most posts with a positive sentiment was Evan Mobley, a member of the Cleveland Cavaliers. It is suspected his posts with a positive sentiment significantly impacted the sentiment of the team.

This type of player specific sentiment analysis could assist advertisers or sponsors identify players who hold a positive sentiment amongst fans. A negative sentiment could help advertisers, sponsors, and agents work to improve the sentiment through a variety of ways such as public relations work. Individuals who play competitive fantasy basketball can use sentiment analysis to determine which players are worth starting or benching on their respective teams. Additionally, team sentiment analysis can be used along with other metrics to form an outlook on performance and assist those who wish to wager on odds. It can also be used by anyone hoping to make media content about popular teams that can generate the most engagement from their supporters.

7 Future Work

7.1 Improvements

If afforded more time, creating a custom lexicon to capture context of the sport would be ideal. Irony and sarcasm are difficult to address, but some basketball related humor could be added to a custom lexicon. Additionally, expanding the targeting of players and teams to include shortened names, abbreviated names, and nicknames would yield more post data to be reviewed, and this expanded dataset could give more insightful results about teams and players.

7.2 Text Summarization

This project could be expanded to include text summarization in order to obtain an accurate summary of what reddit users on the NBA subreddit are posting about teams or players. Additionally, summarization could be used to generate new posts for a target player or team to help shift sentiment in either direction. For instance, results showed Russell Westbrook as the player with the most posts containing a negative sentiment. Only the positive posts can be summarized to shift the sentiment towards a positive one. Thus increasing positive sentiment for the player and the team as a whole.

7.3 Ethical Considerations

This project did not present many ethical implications. The data was collected without any usernames or other identifying information that posed privacy concerns to those posting on the NBA subreddit. NBA players are often subject to careful analysis from a variety of domains, so the sentiment analysis does not intrude any further than what is already publicly shared on reddit.

Those wishing to expand the project can certainly use the results of this data to target players for text summarization as I alluded to as possible future work. This direction could pose ethical implications since sentiment can be artificially boosted in either direction and lead to real world issues. Such issues like driving arguments or validating biases can influence users in the NBA subreddit to hold certain opinions of players or teams. For this reason, the ethical implications of expanding this research should be carefully considered.

Bibliography

[1] Xavier Ferrer Aran et al. "Discovering and Categorising Language Biases in Reddit". In: *CoRR* abs/2008.02754 (2020). arXiv: 2008.02754. URL: https://arxiv.org/abs/2008.02754.

[2] "Enhancing deep learning sentiment analysis with ensemble techniques in social applications". In: *Expert Systems with Applications* 77 (2017), pp. 236–246. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2017.02.002. URL: https://www.sciencedirect.com/science/article/pii/S0957417417300751.

[3] Man Hung et al. "Social network analysis of COVID-19 sentiments: Application of artificial intelligence". In: *Journal of medical Internet research* 22.8 (2020), e22590.

[4] Clayton Hutto and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: 8.1 (2014).

[5] I. Kolyshkina and G. Levin B.and Goldsworthy. "Using Social Media Data for Comparing Brand Awareness, Levels of Consumer Engagement, Public Opinion and Sentiment for Big Four Australian Banks". In: (2013), pp. 59–63.

[6] Kalyan Sahu, Yu Bai, and Yoonsuk Choi. "Supervised Sentiment Analysis of Twitter Handle of President Trump with Data Visualization Technique". In: (Jan. 2020), pp. 0640–0646. DOI: 10.1109/CCWC47524.2020.9031237.

[7] Robert P. Schumaker, A. Tomasz Jarmoszko, and Chester S. Labedz. "Predicting wins and spread in the Premier League using a sentiment analysis of twitter". In: *Decision Support Systems* 88 (2016), pp. 76–84. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2016.05.010. URL: https://www.sciencedirect.com/science/article/pii/S0167923616300835.