Exploring Fan Sentiment in the Digital Realm

Dylan Jorling

May 24, 2023

I. Introduction

In recent years, social media platforms have become a significant source of user-generated content that reflects people's opinions, sentiments, and experiences. Analyzing and understanding the sentiment expressed in these social media posts can provide valuable insights into public opinions, consumer behavior, and market trends. One popular social media platform for studying sentiment analysis is Reddit, a vast online community where users share and discuss various topics.

Social media sentiment analysis involves extracting and analyzing the sentiment expressed in textual data, such as posts, comments, and reviews, to determine overall sentiment polarity (positive, negative, or neutral) and gain a deeper understanding of people's attitudes and opinions. Sentiment analysis has gained considerable attention in the field of natural language processing (NLP) due to its wide range of applications in market research, brand management, public opinion analysis, and customer feedback analysis.

Reddit is a social news aggregation, web content rating, and discussion platform where users can submit posts, comment on them, and engage in conversations. Reddit is structured as a group of subreddits, where each subreddit corresponds to a specific topic. One of the most popular group of subreddits is professional sports, with r/NBA, r/NFL and r/soccer each having several million subscribers. A more granular group of subreddits involves specific team subreddits where fans can interact with each other and discuss everything related to their favorite team. r/Lakers is a subreddit with nearly 500,000 subscribers and is dedicated to the Los Angeles Lakers, a popular National Basketball Association (NBA) professional basketball team.

The purpose of this project is to conduct sentiment analysis over the entirety of the 2022-2023 NBA season using data scraped from r/Lakers. Specifically, the project focuses on an entity-level sentiment analysis related to every player on the team as well as the team owner, general manager and head coach.

Traditional sentiment analysis approaches treat entire text observations as a single unit and assign sentiment scores at the document level. Entity-level sentiment analysis aims to go beyond this by identifying and analyzing sentiment towards specific entities, leading to more complex insights. This paper leverages NLP techniques, including entity recognition, unsupervised sentiment analysis, and data visualization, to gain insights into the sentiment expressed towards each member of the Lakers organization throughout the season. Finally, player sentiment is measured against on-court team and player performance to uncover player-specific relationships.

II. Data Overview

The dataset for this project spans the entirety of the 2022-2023 NBA season from mid-October 2022 to mid-April 2023. To collect the necessary data from Reddit, the Upshift API was used instead of the popular Reddit API which limits scraping to only the most recent 1,000 posts in a specific subreddit.

The final dataset is comprised of full text data for approximately 11,000 posts and 147,000 comments made in r/Lakers throughout the season. The collected data includes essential metadata consisting of post/comment ID, timestamp, author username, upvotes received, and parent post/comment IDs.

III. Text Processing and Methodology

Basic Cleaning and Entity Recognition

Standard text cleaning techniques are applied to the collected data including converting all text to lowercase, removing special characters, eliminating GIFs, and removing hyperlinks. Emojis are deemed important indicators of sentiment and therefore left in the text.

The subsequent task focused on identifying specific players mentioned in each post and comment. Initially, Named Entity Recognition (NER) was employed using a pre-trained model from the SpaCy python library. However, this approach resulted in significant inaccuracies and inconsistencies and as a result, an alternative approach was explored that leveraged both domain knowledge and trial-and-error with regular expression (regex) patterns.

By combining understanding of the entities of interest and iterative testing with regex patterns, a customized approach was developed to identify each specific entity. This approach resulted in initially capturing desired entity references for 76,000 text-mention pairings.

Improving the Base Sentiment Model

Sentiment Analysis was conducted utilizing the Vader Sentiment library in Python, which employs a lexicon-based approach to determine the sentiment of given text. This pre-trained model is specifically trained on social media data and is thus an excellent candidate for use on reddit-sourced data.

Upon testing the sentiment ratings on a subset of several hundred posts and comments, it became evident that the sentiment analysis accuracy was suboptimal. While training a model on a large volume of similar data would provide the most accurate results, such an approach necessitates labeled data in significant quantities and is therefore not realistic. Consequently, incremental adjustments to the lexicon and other easier techniques were identified to enhance performance, resulting in four specific avenues: nickname adjustments, emoji-lexicon adjustments, basketball-related lexicon adjustments, and part-of-speech resolving.

Domain knowledge was leveraged to identify common nicknames associated with each entity that typically implied positive or negative sentiment. These nicknames were extracted and utilized to automatically label posts or comments with respect to the specific entity as either positive or negative. It is important to note that for the numerous posts and comments that mention multiple entities, sentiment is determined with respect to each entity.

Observing the Vader Sentiment Emoji Dictionary revealed that the lexicon ratings for certain emojis were outdated and inaccurate in the context of the data. Additionally, several commonly used emojis were not present in the lexicon at all. To address these issues, missing emojis were added to the lexicon and the sentiment assignments of several other emojis were revised. For example, the sentiment rating for the "fire" emoji initially contained a very negative polarity score, and was thus adjusted to a positive polarity score. In total, 106 emojis were either added to the lexicon or had its sentiment adjusted.

It was determined that incorporating sentiment analysis specific to basketball-related terms could further enhance the model's performance. This involved identifying commonly used basketball-related terms and assigning sentiment polarity scores to them. For instance, a positive sentiment value was assigned to the term "beast" and a negative sentiment value to the term "turnover." Overall, a total 185 basketball-related words were adjusted for.

In specific cases where the sentiment polarity of certain words was context-dependent, SpaCy's part-of-speech (POS) identification was utilized. SpaCy POS attempts to accurately identify the part of speech of each word used in text. One prominent challenge encountered involved the word "like" which has a positive polarity score by default. Upon detailed analysis it was apparent that "like" was frequently used as a preposition, where a neutral sentiment is more appropriate. To address this, text was modified by deleting instances of "like" when used as a preposition, and leaving "like" with a positive polarity score when used as a verb.

These incremental adjustments to the lexicon and the incorporation of domain-specific knowledge aimed to enhance the accuracy and relevance of the sentiment analysis performed on the desired entities in r/Lakers.

Co-reference Resolving, Further Entity Extraction, and Sentence Tokenization

To extract player mentions more comprehensively, a process called co-reference resolving was implemented. Co-reference resolving involves identifying all expressions within the text that refer to the same entity. The utilization of co-reference resolving in this project served two purposes: 1) To extract more intra-text mentions and 2) To extract indirect mentions. Extraction of intra-text mentions involves identifying additional references in text where an entity is already identified and replacing references with the entity name. Applying co-reference resolution to these cases would enhance the viability of using sentence-specific sentiment analysis instead of entire-text sentiment analysis. Extraction of indirect mentions applies to comments that only indirectly mention an entity and made in reply to a parent post/comment with a direct entity mention. Using coreference resolution in these cases resulted in entity identification that otherwise would have been missed.

Co-reference resolution was implemented using the SpaCy-experimental co-reference model, which was trained specifically to identify co-references. The SpaCy NER model was configured to identify actual before the co-reference model identified entity co-references and replaced them with entity names. Due to the aforementioned inconsistencies in SpaCy NER's identification of player entities, custom entities were added to the NER model for all players to ensure their consistent recognition. Since each player was referred to by various names in the data, each mention was replaced with a single name for each entity, substantially reducing the number of custom entities added to the NER model. Furthermore, an entity filter was created to ensure that SpaCy co-reference replace only co-references related to desired player entities, thus drastically speeding up the process.

Following a top-down approach, co-reference resolving was initially applied to all posts containing entity mentions, leading to the extraction of further intra-text entities. Top-level comments, or comments directly replying to a post, were the next set of texts to be resolved. In such cases, the parent-post text was combined with the comment text and the entire text was co-referenced together. This was especially useful in the case where an entity was directly named in a post, but only indirectly named in the comment. The process was repeated for second-level comments by combining the text of the parent first-level comment in a similar fashion and then iterated down five levels of comments. Taking advantage of the hierarchical structure of reddit in this way proved valuable in extracting entity mentions that were missed previously, resulting in an additional 8,000 comment-entity pairings.

The final technique employed to enhance entity-level sentiment accuracy involved implementing sentence tokenization, which involves breaking down text into a list of individual sentences. The rationale behind incorporating sentence tokenization in the sentiment analysis process is to isolate sentences that mention specific entities while disregarding sentences that do not refer to the entity of interest. Intra-text co-reference resolving prior to sentence-tokenization enables all mentions of an entity to be captured and is a much more robust approach than simply utilizing direct mentions.

The evaluation process provides insights into the performance of each method discussed above and contributes to the understanding of entity-specific sentiment within NBA-related discussions on Reddit.

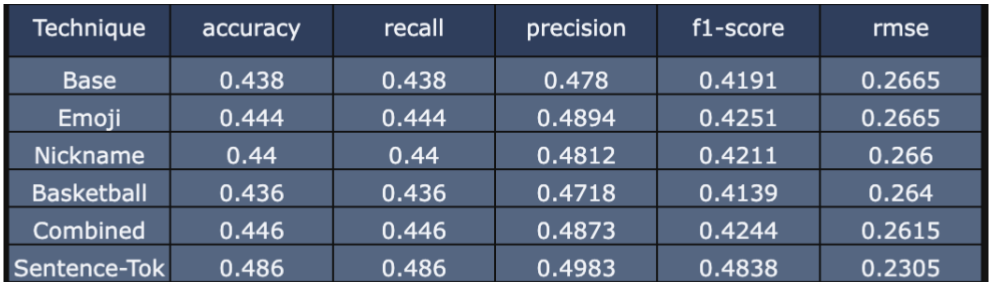
IV. Analysis

Comparing Sentiment Techniques

Sentiment scores were measured for 6 different sentiment techniques:

1) Base: Vader Sentiment lexicon with no adjustments
2) Emoji-Adjusted: Vader Sentiment lexicon with emoji lexicon adjustments
3) Nickname-Adjusted: Vader Sentiment lexicon with no adjustments; positive or negative nicknames automatically assign sentiment
4) Basketball-Lexicon Adjusted: Vader Sentiment lexicon with only basketball-related adjustments
5) Combined: Vader Sentiment with emoji, nickname and basketball adjustments
6) Sentence-Tokenized: Uses combined adjustments and assigns sentiment only for sentences where a specific entity is mentioned

To evaluate the effectiveness of each method, a random sample of 500 text-entity pairings was drawn and manually labeled. The results are shown in Figure 1:



Technique	accuracy	recall	precision	f1-score	rmse
Base	0.438	0.438	0.478	0.4191	0.2665
Emoji	0.444	0.444	0.4894	0.4251	0.2665
Nickname	0.44	0.44	0.4812	0.4211	0.266
Basketball	0.436	0.436	0.4718	0.4139	0.264
Combined	0.446	0.446	0.4873	0.4244	0.2615
Sentence-Tok	0.486	0.486	0.4983	0.4838	0.2305

Figure 1: Comparing Sentiment Performance

Each technique yields marginal improvement in either raw accuracy score, or RMSE, which penalizes two level misses (e.g., positive instead of negative) more than one-level misses (e.g. positive instead of neutral). Combining the first four techniques results in the greatest RMSE reduction and accuracy increase, although the improvement over the base model is marginal.

Utilizing sentence-tokenization resulted in a much more substantial increases in accuracy and f1-score along with a larger decrease in RMSE. While isolating entities using sentence tokenization is extremely useful in determining entity-level sentiment, sentence-tokenization cannot capture intra-sentence sentiment, which would require more advanced techniques. Exploring such techniques in addition to refining the other four used would likely result in even more improvement over the base model.

General and On-Court Performance Analysis

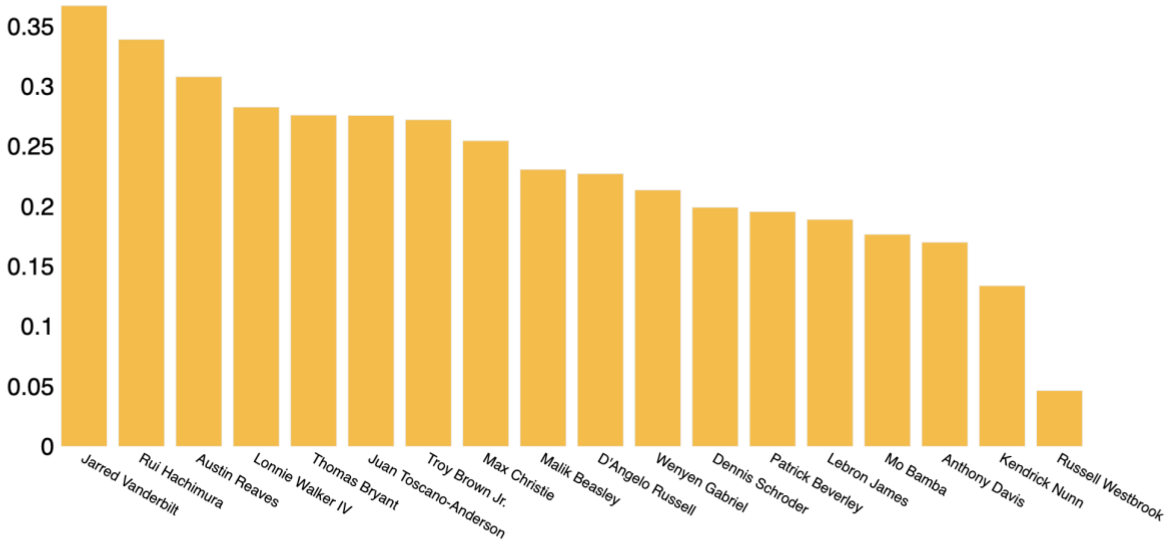
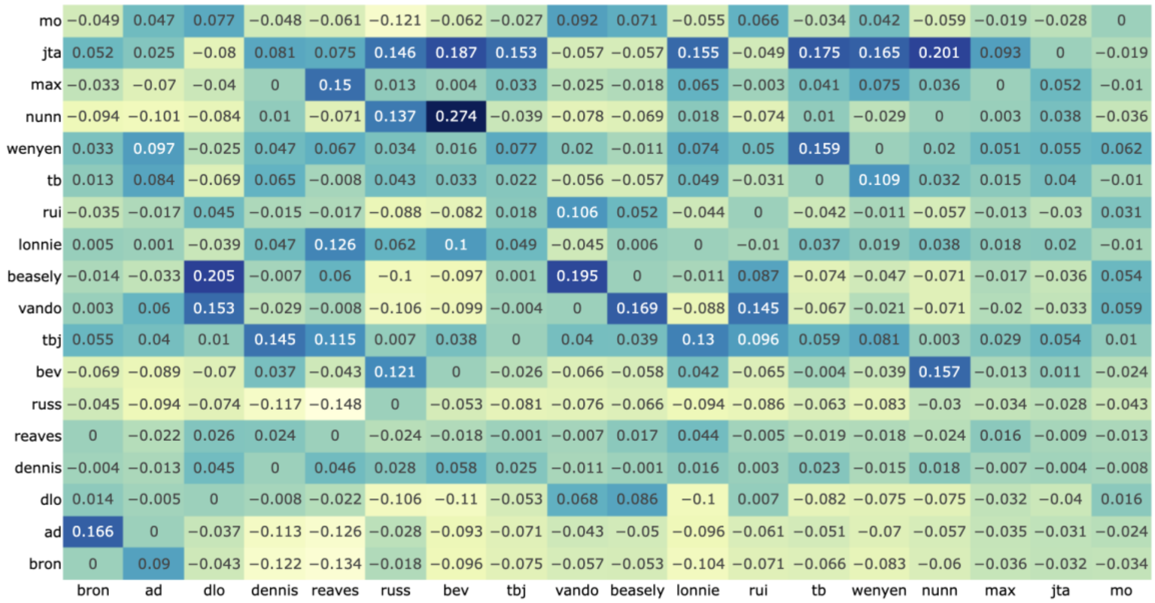


Figure 2: Overall Player Sentiment Scores

Figure 2 displays overall sentiment scores for each player who played significant minutes during the season. Sentiment scores were calculated by subtracting the proportion of negative posts/comments from the proportion of positive posts/comments for each individual player. Additionally, posts were given 2.5x weight compared to comments to account for their higher status. It is unsurprising that the two highest sentiment scores belong to players who were acquired mid-season, given the Lakers' below-average excellent finish. While star players Lebron James and Anthony Davis accounted for approximately 40% of total player mentions, they had two of the lowest sentiment scores on the entire team. Fans were clearly not impressed with the team’s mediocrity for a majority of the season and seemingly put a majority of the blame on its best players.

Young, new players on the other hand had generally higher sentiment ratings than their more established counterparts. This could potentially be linked to excitement over a new, unknown player or enthusiasm about a young player’s potential going forward. Russell Westbrook, acquired the previous season, had a low sentiment score, likely due to the large drop-off in team performance coinciding with his arrival.



	bron	ad	dlo	dennis	reaves	russ	bev	tbj	vando	beasely	lonnie	rui	tb	wenyen	nunn	max	jta	mo
mo	−0.049	0.047	0.077	−0.048	−0.061	−0.121	−0.062	−0.027	0.092	0.071	−0.055	0.066	−0.034	0.042	−0.059	−0.019	−0.028	0
jta	0.052	0.025	−0.08	0.081	0.075	0.146	0.187	0.153	−0.057	−0.057	0.155	−0.049	0.175	0.165	0.201	0.093	0	−0.019
max	−0.033	−0.07	−0.04	0	0.15	0.013	0.004	0.033	−0.025	−0.018	0.065	−0.003	0.041	0.075	0.036	0	0.052	−0.01
nunn	−0.094	−0.101	−0.084	0.01	−0.071	0.137	0.274	−0.039	−0.078	−0.069	0.018	−0.074	0.01	−0.029	0	0.003	0.038	−0.036
wenyen	0.033	0.097	−0.025	0.047	0.067	0.034	0.016	0.077	0.02	−0.011	0.074	0.05	0.159	0	0.02	0.051	0.055	0.062
tb	0.013	0.084	−0.069	0.065	−0.008	0.043	0.033	0.022	−0.056	−0.057	0.049	−0.031	0	0.109	0.032	0.015	0.04	−0.01
rui	−0.035	−0.017	0.045	−0.015	−0.017	−0.088	−0.082	0.018	0.106	0.052	−0.044	0	−0.042	−0.011	−0.057	−0.013	−0.03	0.031
lonnie	0.005	0.001	−0.039	0.047	0.126	0.062	0.1	0.049	−0.045	0.006	0	−0.01	0.037	0.019	0.038	0.018	0.02	−0.01
beasely	−0.014	−0.033	0.205	−0.007	0.06	−0.1	−0.097	0.001	0.195	0	−0.011	0.087	−0.074	−0.047	−0.071	−0.017	−0.036	0.054
vando	0.003	0.06	0.153	−0.029	−0.008	−0.106	−0.099	−0.004	0	0.169	−0.088	0.145	−0.067	−0.021	−0.071	−0.02	−0.033	0.059
tbj	0.055	0.04	0.01	0.145	0.115	0.007	0.038	0	0.04	0.039	0.13	0.096	0.059	0.081	0.003	0.029	0.054	0.01
bev	−0.069	−0.089	−0.07	0.037	−0.043	0.121	0	−0.026	−0.066	−0.058	0.042	−0.065	−0.004	−0.039	0.157	−0.013	0.011	−0.024
russ	−0.045	−0.094	−0.074	−0.117	−0.148	0	−0.053	−0.081	−0.076	−0.066	−0.094	−0.086	−0.063	−0.083	−0.03	−0.034	−0.028	−0.043
reaves	0	−0.022	0.026	0.024	0	−0.024	−0.018	−0.001	−0.007	0.017	0.044	−0.005	−0.019	−0.018	−0.024	0.016	−0.009	−0.013
dennis	−0.004	−0.013	0.045	0	0.046	0.028	0.058	0.025	−0.011	−0.001	0.016	0.003	0.023	−0.015	0.018	−0.007	−0.004	−0.008
dlo	0.014	−0.005	0	−0.008	−0.022	−0.106	−0.11	−0.053	0.068	0.086	−0.1	0.007	−0.082	−0.075	−0.075	−0.032	−0.04	0.016
ad	0.166	0	−0.037	−0.113	−0.126	−0.028	−0.093	−0.071	−0.043	−0.05	−0.096	−0.061	−0.051	−0.07	−0.057	−0.035	−0.031	−0.024
bron	0	0.09	−0.043	−0.122	−0.134	−0.018	−0.096	−0.075	−0.057	−0.053	−0.104	−0.071	−0.066	−0.083	−0.06	−0.036	−0.032	−0.034

Figure 3: Relative Player Associations

Figure 3 displays a heatmap of relative player associations. The plot highlights which players are most often mentioned with each other, relative to how often they are mentioned with all players. For example, the column for titled “bron”, displays the relative percent mentions of Lebron James with every other player. On average, Lebron James is mentioned in 26.1% of posts/comments where another player is mentioned. This 26.1% was subtracted from the initial proportions to adjust for Lebron being an extremely popular player mentioned in a high percentage of posts. The remaining values are thus the proportion of posts/comments above or below 26.1% that Lebron James is mentioned in, for each player.

The plot reveals that star players LeBron James and Anthony Davis are often mentioned together and that player acquired at the same time are also often grouped together. Austin Reaves, considered a "star role player," is much more often mentioned with team role players than team stars. Reaves has one of the highest sentiment scores on the team, and with that context, his low correlations with the star players and their relatively low sentiment scores make sense. Further exploration of these less obvious insights, including breaking down positive and negative mentions, could provide interesting insights into how fans group certain players.

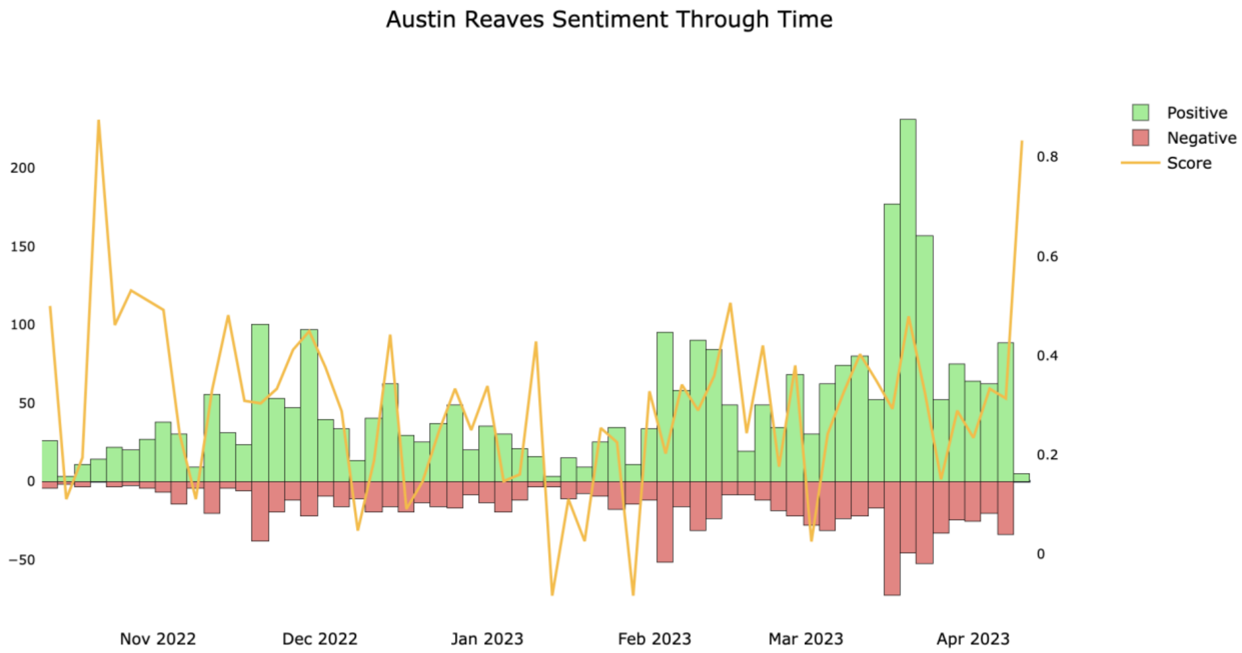


Figure 4: Austin Reaves Sentiment Through Time

Analyzing stats vs player sentiment is another way to learn more about how fans think and whose sentiment is most and least sensitive to both individual and team-based performance. Figure 5 plots the correlation between rolling 20-day player sentiment and rolling 20-day Net Rating. Net Rating is a stat that quantifies a team's point differential per 100 possessions while a certain player is on the floor. For example, if Anthony Davis played 1000 total possessions and the Lakers outscored opponents by 95 points in those possessions, his Net Rating would be +9.5. Net Rating is an easily quantifiable stat that does an adequate job determining how well a player is playing.

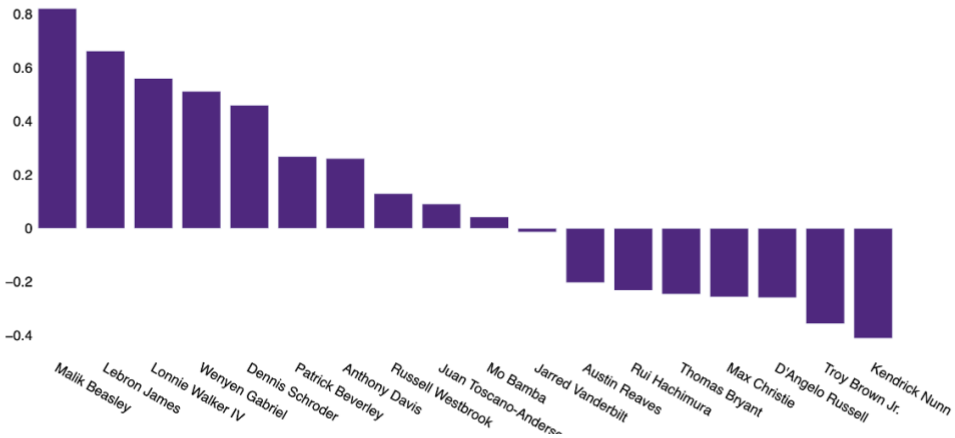


Figure 5: Player Sentiment-Net Rating Correlations

Malik Beasley, a solid yet unspectacular role player, has the highest sentiment-Net Rating correlation on the team. This could potentially be attributed to his high-variance play style as a high-volume, somewhat inconsistent three-point shooter. Additionally, his status as a new player acquired midseason likely contributes to this notion. It is not surprising to see fan sentiment toward star player LeBron James's being highly correlated with his on-court performance, but it is notable that the other team star, Anthony Davis, is more middle-of-the-pack in terms of sentiment correlation. On the other hand, players with negative correlation are mainly young players with future potential and generally high overall sentiment scores. Figure 6 below highlights the stark differences between Malik Beasley and Troy Brown Jr. in terms of sentiment-net rating correlation.

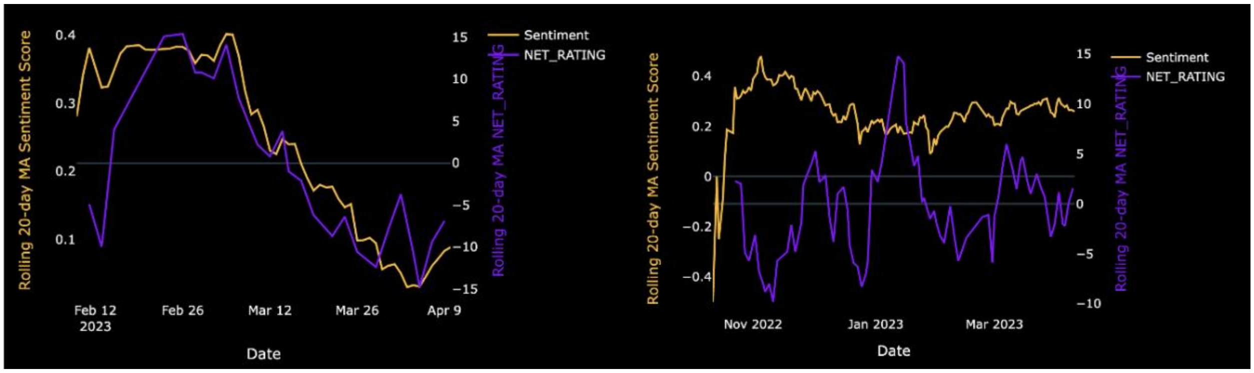


Figure 6: Malik Beasley vs. Troy Brown Jr. Sentiment-Net-Rating

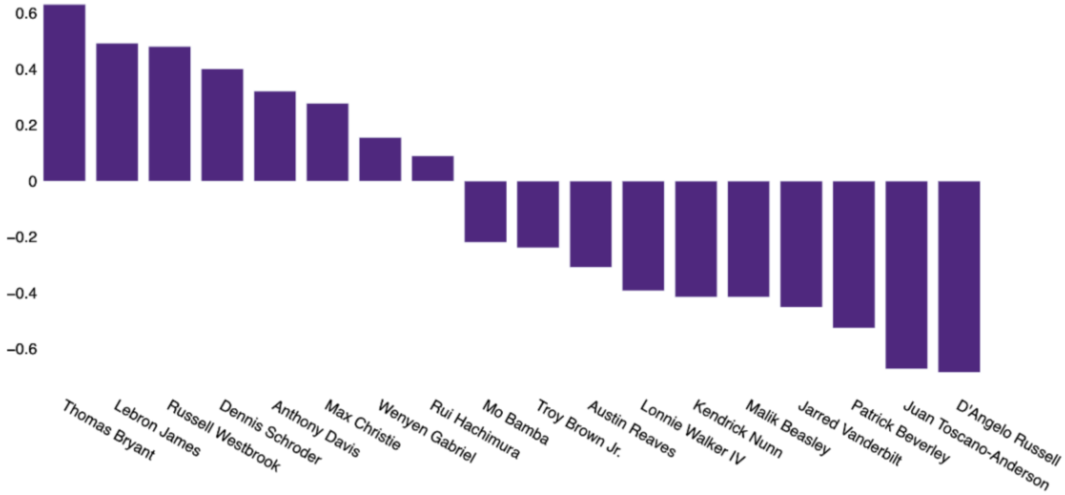


Figure 7: Player Sentiment-Team Wpct Correlations

Moving beyond individual statistics, Figure 7 measures each players' sentiment correlation with team winning percentage. At first glance, the plot appears to be evenly dispersed, with a slight negative skew. It is interesting to note that more players' sentiment is negatively correlated with team winning percentage than positively correlated. Sentiments towards the team's top players are amongst the most positively correlated with team winning percentage. Taken together, this indicates that sentiment towards star players is highly dependent on team performance, while sentiment towards role players is much more related to individual performance. Figure 8 displays two ends of this spectrum: Lebron James and D'Angelo Russell.

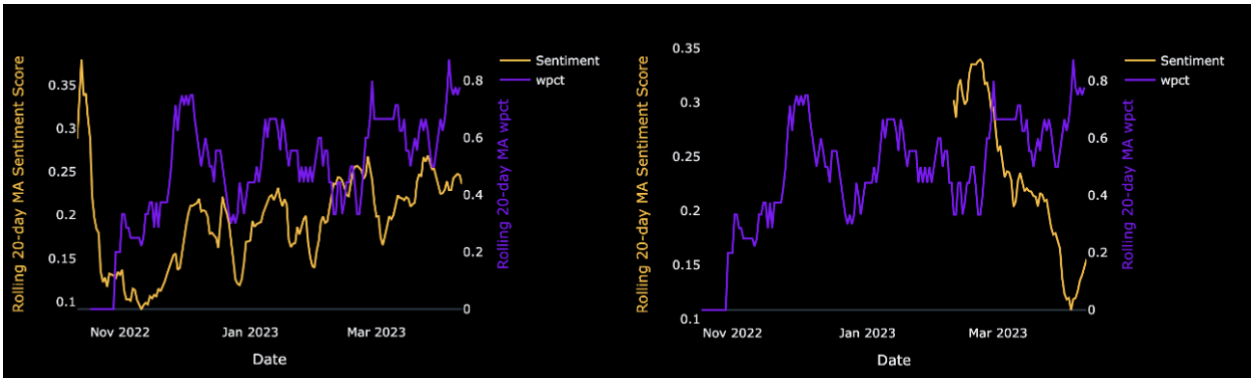


Figure 8: Lebron vs. D'Angelo Russell Sentiment-Team-Wpct

Entity-Level sentiment analysis provides far deeper and less obvious insights compared to traditional sentiment analysis with the examples above being just the tip of the iceberg. This type of analysis can provide valuable information regarding fan behavior and has potential as a key marketing tool.

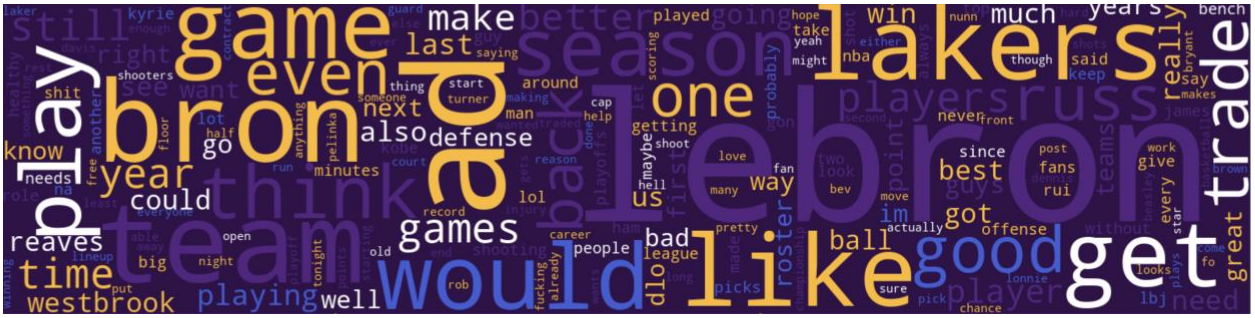
V. Future Research and Conclusion

This sentiment analysis project successfully achieved its objectives of testing and enhancing entity-level sentiment analysis techniques, examining the correlation between entity-level sentiment and basketball statistics, and providing valuable insights into fan behavior. Future research directions may include exploring more mathematically-based lexicon adjustments to further improve sentiment analysis accuracy and developing methods for intra-sentence sentiment detection, allowing for a deeper understanding of sentiment nuances.

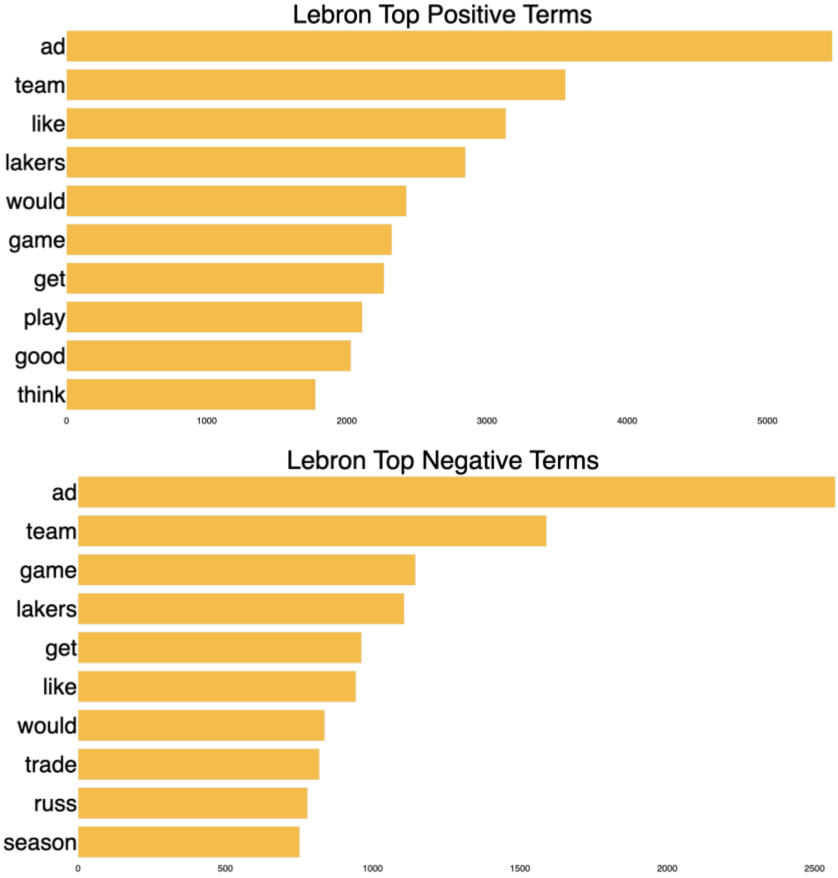
The findings of this project emphasize the power of player-level sentiment analysis in uncovering meaningful patterns in fan behavior, which can be instrumental in shaping effective marketing strategies. Moreover, entity-level sentiment analysis has broad applications beyond the realm of basketball, such as analyzing sentiment towards specific stocks on social media or gauging public sentiment towards potential Presidential candidates. The versatility of this analysis approach makes it a valuable tool in diverse domains and opens up countless opportunities for its application.

In conclusion, this sentiment analysis project has demonstrated the significance of entity-level sentiment analysis techniques and their potential impact on understanding fan sentiment and informing marketing decisions. By continuously refining and expanding these techniques, researchers and practitioners can unlock deeper insights into consumer behavior and sentiment dynamics across various domains, facilitating the development of targeted strategies and improving decision-making processes.

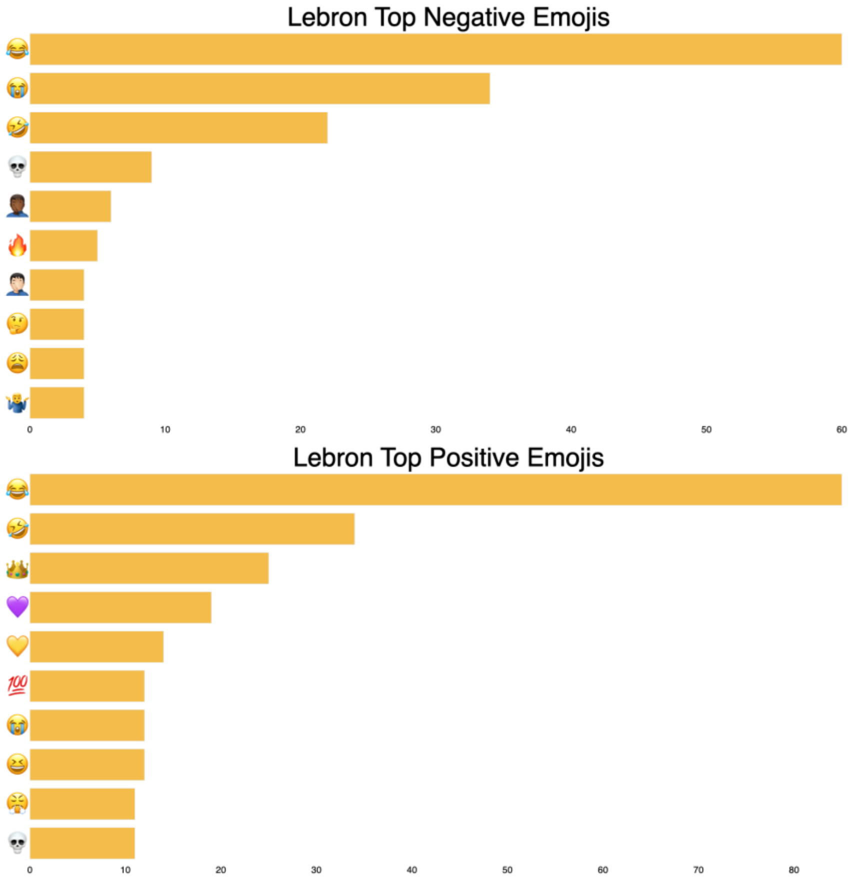
VI. Appendix Plots



Appendix i: Lebron James WordCloud

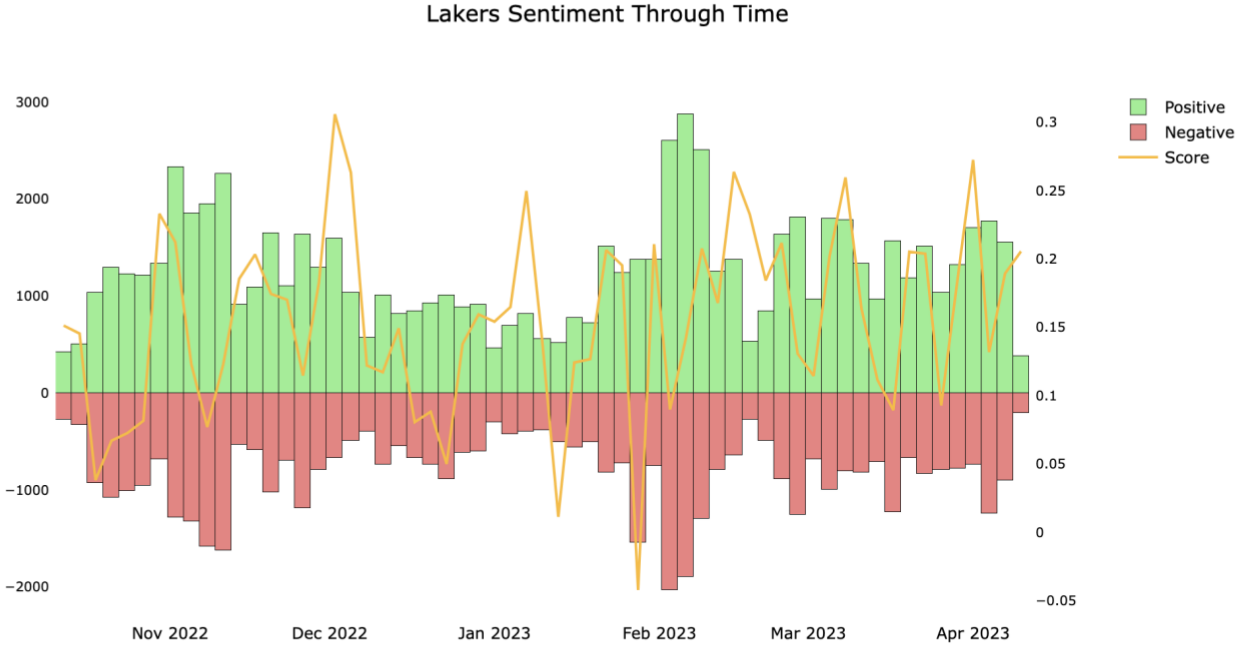


Appendix ii: Lebron James Top Positive/Negative Terms

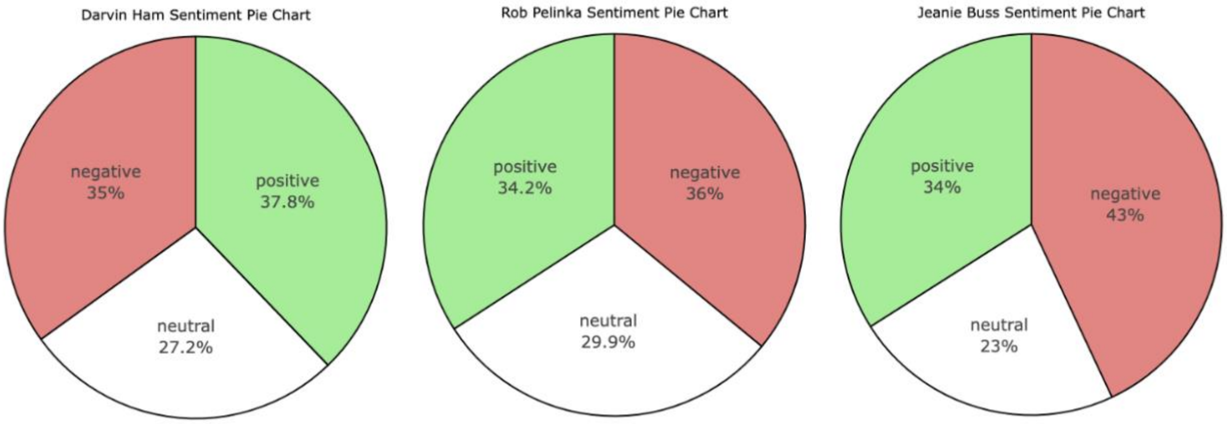

Lebron Top Negative Emojis

Lebron Top Positive Emojis

Appendix iii: Lebron Top Positive/Negative Emojis


Lakers Sentiment Through Time

Appendix iv: r/Lakers Overall Sentiment Through Season



Appendix v: Lakers Coach, GM, Owner Sentiment Pie Charts

`