The voice of the fans: Analyzing Twitter data and generating headlines for the Dallas Mavericks Season

GROUP 4

Pietro Del Bianco (3091355), Ludovico Pasquali (3095581), Antonio Preiti (3104543), Tommaso Renso (3093876), Gaia Triggiani (3110044)

Introduction

The use of social media platforms has become increasingly prevalent in capturing sentiment and opinions on various topics, including sports. In this report we present a comprehensive analysis of the tweets pertaining to the Dallas Mavericks 2021/2022 season, focusing on sentiment analysis and headline generation based on the extracted information. By leveraging the vast amount of user-generated content on Twitter, we aimed to gain valuable insights into the public's perception on the team's performance throughout the season.

The Dallas Mavericks are a prominent professional basketball team in the NBA and have gathered a significant following and fan base. By harnessing the valuable resources provided by Twitter we sought to understand the dynamics of public opinion and how it correlated with the team's on-court performance.

The primary objective of this project was two-fold: first, to perform sentiment analysis on the collected tweets and to spot any weird thing if compared to the team results in the season; and second, to utilize the extracted sentiment information to generate engaging and informative headlines for match summaries. Such insights can prove invaluable to media outlets, sports organizations and fans alike in understanding the overall reception of a team's performance and crafting compelling narratives to engage with the audience.

In this report, we will discuss our data collection and preprocessing methodology, the sentiment analysis techniques used. We will also outline our approach to headline generation, present our overall findings, and provide suggestions for future research and improvement.

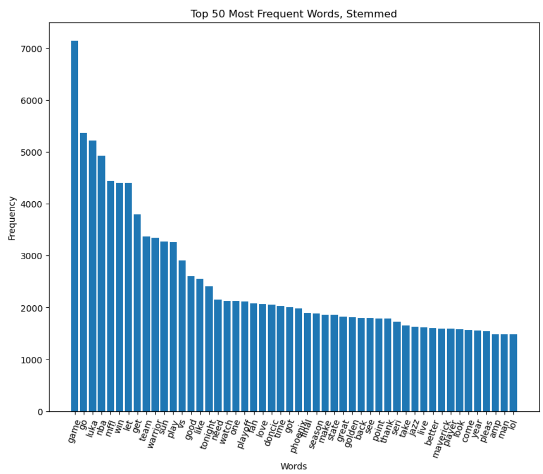
Data collection and preprocessing

Our data collection process began with a dataset obtained from Kaggle, specifically focusing on tweets related to the exceptional performance of the Dallas Mavericks during the

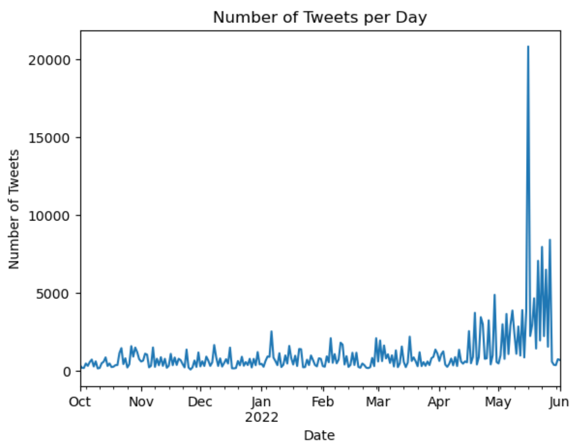
2021/2022 season. From the original dataset of 343,000 tweets, we refined it by filtering out non-English tweets, resulting in a dataset of 263,000 English-language tweets.
The dataset contained valuable features such as datetime, user location, and user-specific details, offering insights into the demographics and characteristics of the Twitter users engaged in discussions about the Mavericks' season.

To prepare the tweets for sentiment analysis, we performed various preprocessing steps. This included removing tags (@) and hyperlinks (‘http’), followed by tokenization to break down the tweets into individual words. Additional techniques such as removing common stop words and applying stemming and lemmatization were utilized to standardize word forms and reduce variations. As a result, we added two new columns to the dataset to store the stemmed and lemmatized versions of the tweets.

To avoid potential bias in sentiment analysis, we excluded Dallas Mavericks-related terms from the exploratory analysis. This ensured that the sentiment analysis focused on the broader sentiment expressed in the tweets rather than being influenced solely by team-specific terms that coil skew our sentiment analysis results by overshadowing other sentiment-bearing words.



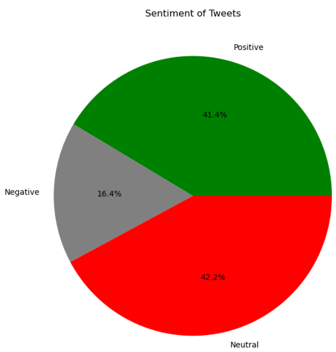
Examining the most frequent unigrams and bigrams provided insights into the key topics and discussions surrounding the Mavericks' season. Terms such as "go", "game", "Luka" (star player Luka Dončić), "MFFL" (Mavericks Fan For Life), "win," "Warriors," and "Suns" (teams faced in the playoff) were prominent, reflecting the enthusiasm and engagement of fans in their conversations.



In addition to analyzing the content of the tweets, we also examined the temporal aspects of the Twitter activity, observing trends in the number of tweets per datetime. Playoff games garnered the highest average tweet count, peaking at around 20,000 tweets, indicating increased engagement and excitement during those crucial matches. In contrast, regular-season games saw fluctuating tweet counts ranging from 1,000 to 5,000, reflecting varying levels of interest and discussion across different games.

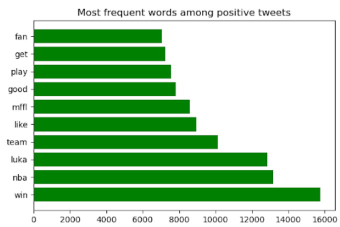
Sentiment Analysis

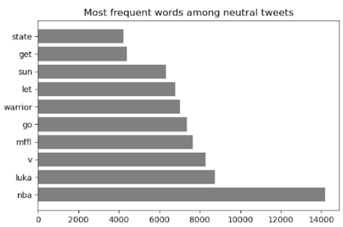
Our sentiment analysis aimed to evaluate the overall sentiment expressed in tweets and its correlation with the Dallas Mavericks' performance during the 2021/2022 season. Initially, we intended to employ RoBERTa, a state-of-the-art language model for sentiment extraction. However, due to computational limitations, we opted for the SentimentIntensityAnalyzer() tool as an alternative approach, which effectively captured the general sentiment of each tweet.

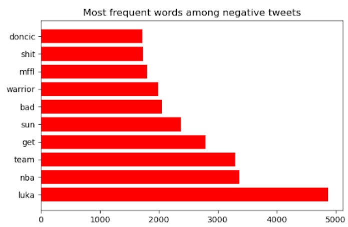


Analyzing the sentiment of the collected tweets revealed a relatively balanced distribution between positive and negative sentiments, with each accounting for approximately 42% of the overall sentiment. This balance indicated that Twitter users expressed a mix of positive and negative opinions about the team's performance. Additionally, around 16% of the tweets fell into the neutral sentiment category, indicating a lack of strong sentiment expression in some cases.

To gain further insights, we delved into the most frequent words found within tweets categorized as positive or negative sentiment. However, we noticed a considerable overlap in the words appearing in both categories. This finding implies that certain terms were commonly used across tweets, regardless of the sentiment expressed. It suggests that the sentiment analysis based on individual words alone may not fully capture the complexity of opinions and emotions shared by Twitter users.







These findings emphasize that sentiment analysis alone may not provide a definitive reflection of the Dallas Mavericks' performance. Sentiments expressed in tweets are influenced by various factors, including real-time game outcomes, player performances, and personal biases of Twitter users. To obtain a comprehensive understanding, it is crucial to contextualize the sentiment analysis within the broader landscape of the team's performance, considering objective measures such as game statistics and win-loss records.

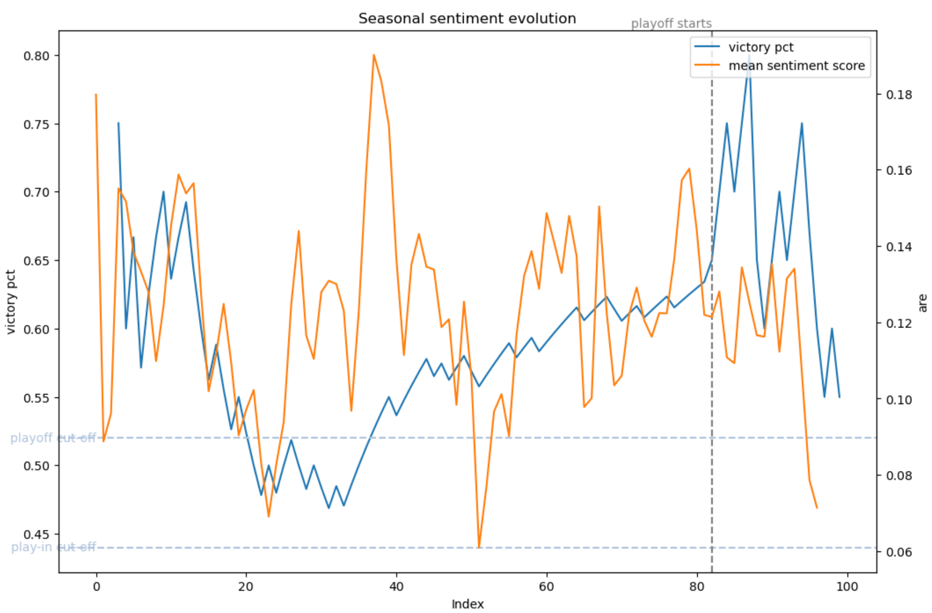
Since our dataset only contains information from tweets scraped from Twitter, we need to devise a performance metric that considers the format of the NBA season. The NBA season consists of two main stages: the Regular Season and the Playoffs. During the Regular Season, teams play 82 matches to determine the top 8 teams from each Conference (Eastern and Western) that qualify for the Playoffs. The Playoff phase begins with the

Quarter Finals (Conference), followed by the Semi Finals (Conference), the Conference Finals, and finally, the NBA Finals. In each series, a team must win 4 out of 7 games to advance to the next round.
Considering this format, our performance metric should account for the different stages of the season.

For the Regular Season, we have chosen a metric called "record" to assess the performance of teams. The record is calculated by dividing the number of matches won by the total number of matches played. Historically a record of 0.44 would lead a team to the Play-in match, which means they would have a chance to qualify for the Playoffs. On the other hand, a Record of 0.55 or higher would secure a team's direct qualification to the Playoffs[1] .

For the playoffs the record metric was computed differently. Once a team reaches the Playoffs, their overall season is considered positive. Therefore, the Record for this phase is computed as follows: $0.5 + \frac{\#\, wins\ in\ the\ series}{4} - \frac{\#\, losses\ in\ the\ series}{4} + 0.1 \times series\ won$
It is important to note that the Record for the Playoff phase tends to be more unstable compared to the first phase. To address this, the Record is normalized to ensure that values always fall between 0.55 and 0.8. Normalization helps provide a consistent range for evaluating performance during the Playoff phase.

Next we defined a sentiment metric by calculating the mean compounded sentiment value for each match, considering tweets published on the match day and the day after until 4 am. However, assessing performance solely based on individual match sentiments would oversimplify the analysis. To address this, we also incorporated tweets from the previous match and assigned them a lower weight by multiplying their compounded sentiment values by a constant decrease factor of 0.45: a single match's performance should not outweigh a series of poor or outstanding performances!



1. NBA League Averages - Per Game | Basketball-Reference.com

The analysis of the results revealed a consistent alignment between sentiment and performance in most instances. When the Dallas Mavericks performed well in their matches, the sentiment expressed in the tweets was predominantly positive, reflecting the fans' satisfaction and excitement. Conversely, during periods of poor performance or unfavorable outcomes, the sentiment tended to be negative, mirroring fans' disappointment and frustration. However, there were a few notable discrepancies that emerged:

- Around game 40, which coincided with January 1st, a peak in sentiment was observed, primarily due to an influx of congratulatory tweets for the new year. While this sentiment surge was unrelated to the team's performance, it highlights the influence of external factors on sentiment analysis.
- A noticeable dip in sentiment occurred around game 54. This decline can be attributed to the trade that involved the Mavericks selling one of their star players, Porzingis, in exchange for Dinwiddie, who is considered a role player rather than a star. This transaction resulted in dissatisfaction among fans, not agreeing with the team's decision.
- Similarly, a low sentiment value was observed around game 72. This dip was a consequence of a substantial loss suffered by the Mavericks against the Minnesota Timberwolves, coupled with a missed opportunity to secure a higher standing in the league as the fourth-ranked team.
- Lastly, towards the end of the season, a decline in sentiment was noted, primarily attributed to the Mavericks' loss in the NBA Finals against the Golden State Warriors. The magnitude of this defeat significantly impacted the sentiment expressed by fans.

These discrepancies reveal the complexity of sentiment analysis in the context of sports: while there is generally a correlation between team performance and sentiment, external events, individual player transactions, and specific game outcomes can also exert a significant influence on the emotions expressed by fans. It underscores the importance of considering various factors and contextual information when interpreting sentiment analysis results to avoid oversimplification and provide a more nuanced understanding of fan sentiments throughout the season.

Headline Generation

The headline generation process involved several steps to ensure informative and relevant newspaper headlines for each match of the Dallas Mavericks season. Initially, we focused our attention on tweets specifically related to each match. This was achieved through a two-step filtering process. Firstly, we selected tweets published on the date of the respective match. Secondly, we performed text classification using a pre-trained NLI-based Zero Shot Text Classification model to distinguish informative tweets labeled as 'match' from the others.

To improve the performance of the headline generation model, we used a cleaned version of the tweets instead of the raw data. To further refine the tweet data, we employed the BART

model, fine-tuned on CNN Daily Mail, to summarize each tweet. This step aimed to normalize the varying styles and lengths of the tweets.
Since the joint version of all the summarized tweets often exceeded the maximum token limit (1024 tokens) accepted by the summarizer model, we partitioned the tweets based on their token count. By iterating through the dataset and tracking the token count, we created partitions of tweets that fell within the acceptable token range. During the joining process, we ensured that each tweet ended with a period to indicate separate concepts and sentences.

Using each partition of the summarized tweets as input, we generated different headlines by setting the maximum length parameter of the model to an average number of tokens for headlines, such as 20. This approach allowed us to obtain diverse and concise newspaper headlines for each match, capturing the key information and essence of the game.
Once we obtained different partitions for every game we implemented the summarizer again to obtain one single and comprehensive headline for every game.

For the sake of the analysis and for computational reasons we focused on a selected subset of games from the Dallas Mavericks' 2021/2022 season, specifically highlighting those that were particularly intriguing and noteworthy. This subset included the early games where star players delivered outstanding performances, as well as games featuring dramatic losses. Additionally, we included some of the most thrilling playoff games, such as the epic Game 7 against the Suns, which propelled the Mavericks to the conference finals.

Baseline

Before proceeding with the evaluation of the model, we established a basic baseline approach to further assess the effectiveness of the more complex headline generation method described earlier. The defined baseline is straightforward and involves generating one of two potential headlines ('Dallas Mavericks won/lost the match') based solely on the perceived sentiment of the tweets related to that particular match, without considering any additional news or information derived from the tweet texts.

Model evaluation

During the model evaluation process, we aimed to assess the performance of our headline generation model by comparing the generated headlines with real headlines extracted from reputable sports news sources such as Sky Sports, Fox Sports, and other NBA newspapers in the USA. We gathered a total of 99 real headlines, one for each game of the season, to provide a diverse and comprehensive set of reference headlines for evaluation.

To measure the quality and similarity of the generated headlines, we utilized the ROUGE score, which is a widely used metric in natural language processing tasks, including text summarization. The ROUGE score evaluates the overlap between the generated headline and the reference headline based on various measures such as n-gram matching, word sequence similarity, and sentence-level recall. It provides a quantitative assessment of how well the generated headline captures the key information and sentiment of the game.

During our evaluation, we achieved a maximum ROUGE score of 0.6 for a game where the generated headline accurately portrayed the Dallas Mavericks' loss against the Los Angeles Clippers, including reporting the final score of 97-91. This demonstrates the model's capability to capture the essential outcome of the game and present it in a concise headline format.

The majority of the ROUGE scores obtained were concentrated in the range of 0.1 to 0.35. In these cases, the generated headlines correctly predicted the final outcome of the game but may have missed some additional details or nuances. For instance, the model might not have included information about winning streaks, series of negative performances, or the comeback of previously injured players. This limitation could be attributed to the scarcity of tweets discussing these specific aspects, as the model primarily relied on the available match-related tweets for generating headlines.

However, there were instances where unexpected events occurred during the games. In some cases, the model successfully captured and reported these events, such as a late overtime win against the Clippers, which is quite exceptional in NBA games. However, there were other events that the model failed to include in the generated headlines, such as a game delay due to a leaking roof. This discrepancy could be attributed to the model's reliance on the information present in the match-related tweets, which might not have adequately captured certain occurrences or details.

Conclusion & future works

Overall, the model evaluation results indicate a satisfactory performance in generating accurate headlines that align with the final outcomes of the games. While the model demonstrated the ability to capture the core information and results, there is room for improvement in incorporating additional contextual details and unexpected events. Future enhancements could involve expanding the data sources and incorporating a wider range of social media platforms to capture a more comprehensive view of the games and the associated discussions, enabling the model to generate more informative and contextually rich headlines.