Allegheny College
Computer And Information Science Department

Senior Thesis

Hoop Hype:Measuring Media Sentiment to Quantify NBA Team Perception

by

Luke Barker

Allegheny College

DEPARTMENT OF COMPUTER AND INFORMATION SCIENCE

Project Supervisor: **Janyl Jumadinova**
Co-Supervisor: **Emily Graber**

Abstract

The motivation behind my project is the way that the current NBA media space is right now. This project aims to expose media bias in NBA Media. Studies have been done performing sentiment analysis using combining different variations of methods, but none using SpaCyTextblob in articles about NBA media. Also, this project is the first of its kind to look at media bias specifically in the NBA and display the results on a deployed dashboard. The data for this project was accumulated over multiple months during the 2024-2025 NBA season using articles from NewsAPI with the query search being set to “NBA”. Once the articles were collected, I used a tool called SpaCyTextblob by SpaCy to perform the sentiment analysis. The dashboard displaying the final results was made using a tool called 11ty which can create a simple HTML dashboard for a static site. I tested other sentiment analysis tools to make sure that I chose the correct one for sports related media articles. The second experiment that I ran was testing different sized data files and seeing which file size of data would produce the best results. This project is a building block for exposing media bias in the NBA. One direction that this project could be furthered in would be analyzing the sentiment results and comparing it to real life trends in the NBA and cross reference those trends with my results to see if there is any correlation. Overall this project is a large step towards exposing media bias NBA and is the foundation for future work in this specific area.

Table of contents

List of Figures

List of Tables

1 Introduction

Sports, basketball specifically, represent one of the largest industries in the world, going beyond geographic, cultural, and economic boundaries. From local communities to global arenas, sports bring people together and create moments of shared excitement and passion. The global sports market was generating a lot of money in recent years, with major leagues like the NBA contributing significantly to this figure. The NBA, specifically, generates over $10 billion annually through a combination of ticket sales, broadcasting rights, sponsorships, and merchandise[22]. Basketball, as a global sport, has a massive following, with the NBA as basketball's main attraction and main money-maker. The league's reach extends far beyond the United States, thanks to its international players, global marketing strategies, and the immense popularity of stars who resonate with fans worldwide. Social media and 24/7 news cycles have only made the visibility and impact of sports, with millions of articles, videos, and posts covering games, players, and teams larger.

1.1 Motivation

In today's world with media being the forefront of everything, media narratives hold significant power. They shape how fans perceive teams and players, influence sponsorship deals, and even affect athletes' mental health and career trajectories. With such a vast and influential market, understanding how media operates within the sports industry becomes essential for maintaining fairness and equity.One large issue in the NBA is media bias. With the rise of social media and media itself, it can be easy to be biased towards a team or a player. Media bias can be hurtful for the NBA itself and/or individual players and teams. It can create false narratives about a team or player which could end up hurting a players career or giving an organization in the NBA a bad reputation. The central question being addressed in this project is whether media bias exists within the National Basketball Association (NBA). This is a very important research question and by figuring out the answer to this research question, it could change the way the media looks. This is also an important question because it can lead to potentially helping players careers or helping organizations get more coverage. By conducting this research and finally showing the truth about social media/ media bias in the NBA, people around the world that follow the NBA will now be able to see what has been happening for a long time. This study is important because it allows fans of the NBA that might not have even known that this was happening to now know and see that there is media bias within sports, specifically the NBA. It should enlighten people on how the media can twist people's minds and will get us to think or see whatever they want us to see.

1.2 Goals of the Project

This project,Hoop Hype, centers around the analysis of National Basketball Association(NBA) team mentions from NewsAPI articles, using Natural Language Processing (NLP) techniques and sentiment analysis to show sentiment of NBA teams based on the text in the articles. The main goal of this project is to shine a light on media bias by performing sentiment analysis and popularity analysis on articles that mention NBA teams and NBA players.The main goal of this project is to shine light on this issue by displaying information on a dashboard that shows the difference in how teams and players are portrayed in the NBA in the media. Information on the dashboard is going to display unbiased information about how teams are mentioned at a different rate, and also that they are talked about in certain ways based on what team they are and what players are on that team.

1.2.1 Key Terms

To better understand this study, it is important to define the key terms used throughout the research project. Media bias refers to the inclination or prejudice for or against a particular person, group, or organization in the media. In the context of this project, media bias is examined through the lens of news articles and how they portray NBA teams and players. Sentiment analysis is a Natural Language Processing (NLP) technique used to analyze text data to determine the emotional tone—positive, negative, or neutral. This is important when trying to decide if the media is biased or not. NewsAPI is a tool that takes news content from various sources, serving as the primary source of data for this analysis. Dashboard in this context refers to the visual interface used to present findings, allowing users to compare and contrast team mentions and sentiment scores effectively. The NBA or National Basketball Association is a professional basketball league with 29 teams in the United States and 1 in Canada. These definitions are very important as they are the backbones of the project and without understanding these key terms, it will be difficult to understand how the project will work.

1.2.2 Assumptions / Limitations

This project is based on a few basic assumptions, one being that the sentiment analysis tool works well enough to figure out if something is positive, negative, or neutral, even though it might mess up on certain things like sarcasm. We're also assuming that the articles we're pulling from NewsAPI give a good idea of what's out there in the media about the NBA and are a good representative of the whole league. Another big assumption is that people care enough about this kind of analysis to see how teams and players are being talked about. Basically, we're trusting the tools and the data to give us something useful, even though they're not perfect. There are a few things that might make this project less perfect or limit this project. For one, the sentiment analysis tool isn't always great at figuring out stuff like sarcasm or subtle language, so it might not always get the tone right. Another limitation is that the articles we are looking at are only from one place and it might not cover all of the media about the NBA. For delimitations, we're only focusing on NBA teams and players, not sports in general, and we're sticking to articles from the last five years. This keeps things simple and manageable but means we're not looking at the whole picture.

1.3 Ethical Implcations

Several ethical dilemmas could be talked about when conducting this study. One of the main ethical concerns in this project is the potential bias when the sentiment analysis tools are being used. These tools are created using algorithms that learn from large datasets, which means any bias in the training data can carry over into the analysis. For example, if the algorithm was trained on articles that tend to portray certain teams or players negatively, it might reflect that bias in its output, even if the actual sentiment in the articles being analyzed is neutral or positive. This creates a significant risk of misrepresentation in the findings. Additionally, sentiment analysis tools can struggle with things such as sarcasm, metaphors, or complex sentence structures, which are common in sports journalism. These limitations mean the tool might incorrectly classify an important but funny comment as positive or fail to detect the small amount of sarcastic negativity in a sentence. This could lead to skewed results that don't accurately represent how NBA teams and players are portrayed in the media. To try to remove this bias as much as possible requires careful validation of the sentiment analysis tool and acknowledgment of its limitations, but it's still a significant ethical challenge that could impact the credibility of the study.

This project involves analyzing and potentially spotlighting specific media outlets for their coverage of NBA teams and players. If the study reveals significant bias in the articles from certain outlets, it could harm their reputation, especially if readers interpret the findings as a direct critique of those outlets' professionalism or ethics. This could be very damaging if the study's results are publicly shared without proper context or explanation of why they are being used. Media/News organizations might feel like they are being targeted for no reason when in reality they were just reporting on sports like they were supposed to be and not spreading bias. For example, an outlet might cover one team more frequently because it has a larger fan base, not because of bias. It's important to present the findings responsibly to avoid placing unearned blame on any particular outlet, but this is a delicate balance that's not always easy to achieve. This is one of the most important ethical issues to deal with because you could be directly hurting someone. Data that is being used from articles that are open to the public need to be checked to ensure that the people that directly relate to it do not get hurt.

Another ethical dilemma is the possibility that the public might misinterpret the results of this project. The dashboard will present data on media mentions and sentiment, but without context, people might come to a different conclusion. For example, if a team appears to receive more negative coverage, fans or stakeholders might assume that the media does not like that team and they are being biased. On the other hand, the findings could just be the way they are because of things like poor performance or controversies involving the team. Similarly, if a player is shown to have mostly positive coverage, it might not mean they are being favored for no reason. It could just be because they are being successful on the basketball court or something off the court leads to their rise in popularity. The data needs to be shown in a way where it is not misinterpreted. This could stop problems from arising between fans, players, or people in the organization.

The project involves collecting and analyzing a large number of articles from different sources, but there's a chance that certain articles might accidentally stand out. If an article with particularly strong sentiment, whether that be positive or negative, gets a lot of attention in the dataset, it might change the results of the project. This could accidentally put the publisher of the article in the spotlight, which could lead to unwanted criticism or hate. For example, if a single article is unusually critical of a team, viewers of the dashboard might assume that the outlet or author has a bias, even if it's just one article among many. This could harm the reputation of individual writers or news outlets. To not allow this to happen, the project should rely on the patterns of the data and not individual articles and what they say. This would lead to individual writers and news outlets to be protected.

Finally, there's an ethical challenge of presenting the data in a way that is both clear and accurate. The dashboard is to provide an unbiased view of how NBA teams and players are portrayed in the media, but the way the data is visualized could still introduce bias or confusion without explanation. For example, if one team has significantly more mentions than others, it might appear that the media is unfairly focusing on them, even if those mentions are because of the team's recent success or problems. Similarly, if sentiment scores are displayed without context, viewers might not understand the reasons behind them. For example, a neutral sentiment score doesn't always mean the coverage is balanced. The coverage could be very positive or very negative, which would lead to the sentiment score to be neutral. Making sure that the dashboard is user-friendly and that the data is very well explained is very important, but there's still a chance that people will misinterpret the findings or use them to see what they want to see from the results. [9] Building on the ideas presented in the abstract from Luciano Floridi and colleagues, it becomes evident that the ethical responsibility of data presentation is a main component of data ethics. As data ethics emphasizes the moralities of data use and preprocessing, the design and implementation of a dashboard that visualizes media sentiment about NBA teams must also follow these principles. This article highlights the need for a macroethical approach or an approach that recognizes the broader societal implications of how data is used and interpreted. This applies directly to the challenge of visualizing sentiment analysis and media mentions in a sports context, where bias can be unintentionally introduced through design choices or a lack of contextual information that is being presented from the data.

2 Related work

2.1 Sentiment analysis

Sentiment analysis is an important area of Natural Language Processing (NLP) that focuses on identifying and showing the emotions that are in the text provided. It has been applied in many ways before, including running sentiment analysis on customer reviews, political opinions, and social media trends. For this project, sentiment analysis provides a way to evaluate how the media portrays NBA teams and players. By analyzing the tone of the articles which mention players or teams whether that is positive, negative, or neutral, sentiment analysis says a lot about potential biases and trends in media coverage. Modern techniques of performing sentiment analysis allow for more understanding by capturing the context in which words are used. However, sentiment analysis faces significant challenges, such as handling things like sarcasm, which is very common in sports journalism. This makes sentiment analysis both a very powerful tool for this project and a potential limitation in the project scope.

Sentiment analysis significantly influences the project's ability to measure and interpret media narratives. The main part of sentiment analysis in this project is that it is a systematic approach to get the tone of the articles mentioning NBA teams and players.. By using modern NLP techniques like machine learning, the project can process large datasets efficiently, offering insights to the tone of the articles that would be too hard to achieve manually. Sentiment analysis helps uncover patterns in media portrayals, which will then allow people to see potential biases or trends that may not be immediately visible. However, its limitations also shape the project's outcomes. By not being able to detect certain things like sarcasm, when completing the project, it is important to keep in mind that things can be misunderstood without context to the article. Additionally, sentiment analysis may oversimplify the tone of articles by categorizing them as positive, negative, or neutral, missing small things that humans would capture. Despite these challenges, sentiment analysis remains a core component of the project, providing an essential way to look at media narratives about the NBA and be able to analyze and understand them. Sentiment analysis has been used in a large number of projects before.

To get a greater understanding of how it works, it is important to look at how it has been used in the past and in other projects. This also allows you to measure the way it is implemented in this project to the way other people talk about it. An example of sentiment analysis applied to real-world social media data, is shown in the work of Baucom et al. [4], which is a project that shows the potential of combining these two things to gain insights into public opinion. The project uses sentiment analysis to gain insight into whether or not people talk about an NBA team a certain way based on their geolocation. To do this, Twitter API was used to grab tweets as data and then sentiment analysis was performed on the tweets. This project is similar in the way that it uses real world data and performs sentiment analysis on it to try and find something out about, in my case the media bias in the NBA. Chen et al. [5] explored the use of deep learning techniques for sentiment analysis on social media, to try and create their own framework and dictionary for sentiment analysis. This project uses the Spacy natural language toolkit(NLTK) to perform sentiment analysis and this article describes and gives insight on how sentiment analysis is classified. Feddersen et al. [8] looked at sentiment analysis in sports betting and how that can influence people to bet based on what their favorite team is. Feddersens method to get his data was looking at things related to the NBA such as all star votes and arena capacity. This was his way of evaluating the team's sentiment and then from there he used sports betting as his method of proving the bias. My method was similar in the way that I used sentiment analysis to prove bias in the media in the NBA instead of the sport betting market.

With sentiment analysis being a vital part of my project, it is important to understand how performing sentiment analysis on a piece of text works. Li and Wu [14] demonstrate how text mining and sentiment analysis can be utilized by describing how in their method, each piece of text was given a value based on the overall influence of the text and that is a way to help perform the sentiment analysis. By reading this article, it gave me insight into how the lower level classification works in the sentiment analysis process. Drus and Khalid [7] go over the different methods of sentiment analysis that can be used in social media. The research evaluates the Lexicon based approach and the machine learning approach and which one works better for social media posts. This is important to this research because we are performing sentiment analysis on media and it is important to know which way of performing it is the best and how that way works. Balahur [3] does a very good job at exploring sentiment analysis and going into the methodologies involved in performing the sentiment analysis on media. There are alot of preprocessing steps that are behind the scenes when performing sentiment analysis and it is important to know how these techniques work. For example, when performing sentiment analysis on media, there can be a lot of slang and weird use of capitalization. To take care of this, in the sentiment analysis process, there are things that go on behind the scenes such are tokenization, lower casing of text, and slang replacement. This makes it easier to perform the sentiment analysis and makes the text look more normal and easily readable.

Lastly, when dealing with sentiment analysis, it is important to make sure that the method of sentiment analysis that we are using will actually work. Wunderlich and Memmert [23] highlights the use of lexicon-based sentiment analysis to look at sports related media on twitter to see if it is viable enough to research. This study dives into sentiment analysis and tests how accurate it can be when being performed on social media or media posts about sports. The study took 1000 tweets that were realistic and performed sentiment analysis on them and found out that the lexicon based approach to sentiment analysis was more than 95% accurate. With this being a large part of this research project, it is important the sentiment analysis is performed with accuracy or it could unintentionally discredit the research. There have been many projects that use similar techniques and methods, but none that use sentiment analysis to try and prove media bias in sports.

2.2 Media Bias

Media bias has been studied a lot across different areas such as communication, journalism, and sociology. It refers to the tendency of the media to favor particular perspectives, groups, or agendas because of what they believe in. In sports media, bias shows up in how teams and players are talked about, often based on things like how big a team's market is, how famous its players are, or what fans like. This project looks at media bias in the NBA, where the way teams and players are covered can change how people see them and even affect other things like endorsement deals or the way they make money. Media can do good things like bringing fans together, celebrating achievements, and growing the sport's popularity, but bias can cause problems. This can give more attention to certain teams or players while ignoring smaller teams or ones that aren't doing well. Sometimes, bias can actually help by giving lesser-known teams or players more attention. Because of this mix of good and bad, studying media bias is important and not always simple. This project will help explain how media bias affects NBA teams and players and show both the helpful and harmful sides of it. Media bias is a large part of this project and it is a very large issue in today's society. Rodrigo-Ginés et al. [18] provides a review on media bias detection, which goes over different ways to examine it and review it. The research goes over why people are biased in the media and that there are two main reasons why people are biased. One of them being confirmation bias and the other one being naïve realism. Confirmation bias is believing information based on our existing beliefs, and naïve realism is believing your own conclusions are right which leads to you believing other peoples opinions are wrong or they are misinformed. Both of these are reasons that media bias could be happening and it is important to know the background of this for this research project.

Another thing about media bias is that the people that are consuming the media might not even realize that it is biased because of their beliefs. Cramer [6] conducted a study showing that people chose to expose themselves to media in Wisconsin, which led them to have perceptions about media bias. Understanding whether or not something is media biased or if the viewers just think it is media bias because of their own views is important. Media bias can have a large impact on the way people view certain things. Aggarwal et al. [1] looks at different methods for detecting media bias but also looks at its short-term impact, offering insights into how media bias and people reporting things in a biased way can influence public perception and behavior. This deals with this research because without understanding the outcome that media bias can have on people and society, it is hard to fully understand what the point of detecting media bias is.

Another angle to look at is how it affects the media outlets itself. This is important to take into consideration when doing research that involves trying to use media outlets to prove media bias. Shultziner and Stukalin [21] explore how partisan media bias influences the production of news and the media outlets themselves. Media bias can change how people see a news outlet. If the outlet seems like it is only talking about one particular thing, some people might like it more, but others might stop trusting it. Media bias also makes outlets pick stories that match what their audience wants to hear. Sometimes outlets try to stand out by being more biased, which can make people argue more about the media.

Lastly, when talking about media bias, when it comes to detecting it, the faster the better. For this research it is important to automate this process and make sure that it is being detected in the fastest way possible. Hamborg et al. [10] discusses how automated computer science tools can detect media bias in new articles. In their research, they come to the conclusion that automated methods of finding media bias are going to be more effective and applicable. In the article, it also says that automation will make things more efficient which will benefit the media consumer because they will be able to identify the bias quicker. Media bias is very essential to this project and it is important to look at it from all different angles and be able to understand how it happens, why it happens, and what it means.

2.3 Dashboard

Dashboards play an important role in the success of this project by providing an easy to read and interactive interface for visualizing the data results. In this project, the dashboard's main purpose is to make the information and results that the project shows easy to read and understand which allows the users to explore and interact with the data on their own. Dashboards are important in sentiment analysis and media bias exploration because they offer a clear, visually appealing way to present data that might on the other hand be too confusing to understand or access. Through interactive visualizations and components, the dashboard allows users to engage with data in an easier way. In this project, the dashboard is not just a way to view the data but a vital aspect of the research project because it directly affects how people looking at the research can interpret and view the data.

When deciding how to build and create a dashboard, it is important to think about all of the different ways you can do so and which one best fits the research at hand. Sedrakyan et al. [20] talks about how the dashboard should be set up in a way where the person looking at it is able to learn from the dashboard. This design or setup of a dashboard is the best way to overlay the information to the person interacting with the dashboard. If the person can learn and follow along with the dashboard in this manner, it will better get the results across to them and it will also help them understand it better. Dashboard's have many real applications and can hold important information that can help people. Franklin et al. [12] highlight how dashboard visualizations can make decision making processes quicker by presenting the data through a dashboard efficiently. Their research was to try and test different ways to get information to the emergency department(ED) so that the people working at the hospital get real time information as quickly as possible. This is a real life application of dashboards and it is an example of how important dashboards can be for getting information across when needed. If dashboards are a way to get information to hospital workers in the best and quickest way possible, that ensures that it is a good method of explaining data to people in the best way possible.

Schulze et al. [19] provide a review of digital dashboards used for visualizing public health data. This study provides a review of digital dashboards used for visualizing public health data. When data needs to be visualized in the health field, this study shows that the best method for it is dashboards. This can relate to this research because after the data is done being collected and processed there needs to be a way that it can be visualized. After reviewing this research, it is clear that a dashboard is the best way to display the results from the sentiment analysis.

Ramly et al. [17] did a research project that looks at how different data visualization methods improve operations dashboards, which is a dashboard that deals with business methods, focusing on making information clear and easy to use. Their findings are helpful for this research, which aims to create a dashboard for analyzing media bias in sports. The study highlights the importance of choosing the right charts, layouts, and features to make data easy to understand. By applying these principles, the dashboard in this project can be more user-friendly, helping both experts and casual users find patterns in media coverage quickly and effectively. There are 5 different criteria which are different areas of a dashboard that the research looks at and this is how the best dashboard method is selected for operational dashboards. When completing the dashboard for this research this was very helpful in looking at what makes an interactive and informative dashboard.

3 Method of approach

3.1 Data

One dataset used in this study consists of a CSV file containing a large list of NBA players and the team that they play for. This dataset was collected from a source called kaggle and altered to only have what is needed for this project. Jasin [11] provides a dataset on NBA 2K25 player attributes which was helpful to me because it had every player in the NBA in it, and also what team they played for. I deleted all of the other rows of information since there was no need for them in the completion of this project. This dataset serves as a way for being able to check if the teams or players are mentioned in the sports articles. The CSV file includes columns for player names and the team that they are affiliated with. Mentions are tallied and sentiment analysis is performed on these sports articles, and the dataset is utilized to detect and extract mentions and polarity scores. This process enables an evaluation of how teams and players are portrayed in the media. To preprocess the data before performing sentiment analysis and mention counting, I first ensure that the articles retrieved from NewsAPI are properly structured and contain the necessary content. Since some articles may have missing or empty content fields, I use conditional logic to ignore any entries where the content is unavailable. I then convert all text to lowercase to enable case-insensitive matching when identifying mentions of NBA teams.

3.1.1 Sample of NBA Players Dataset

Name	Team
LeBron James	Lakers
Stephen Curry	Warriors
Giannis Antetokounmpo	Bucks
Jayson Tatum	Celtics
Kevin Durant	Suns

3.2 NewsAPI

The sports articles that I am retrieving are from NewsAPI [16]. NewsAPI is an API where you can pull new articles from various sources for all types of projects. For my project, I used NewsAPI to retrieve articles about the NBA by setting the query to “NBA,” ensuring that only articles mentioning the NBA were collected. The data that I collect from NewsAPI is very important because it allows me to be able to cross-check my CSV data set with the articles to see if the NBA teams/players are mentioned. Both of these datasets are important in being able to perform sentiment analysis and a mention count on NBA teams. My method for collecting the articles from NewsAPI was to collect the articles at the first of every month so that I can have articles about the NBA throughout the season instead of just from one period. This will be beneficial because it will allow me to have a larger overall dataset which will make the results more fair and give each team a fair chance to be represented. The JSON file that has all of the articles is 398 KB large and consists of approximately 328 articles. This is important information because it allows you to be able understand the size of the whole project. To be able to use NewsAPI, you can make a free account and use the API key that they give you in your code. Below is a code snippet of how I implemented it in my code. I create a function that is called get_nba_news that has three parameters. I set the query to NBA like I explained above to get articles about the NBA. The other paremeter is my API key that I got after creating my NewsAPI account. The function will add the articles to a file called data.json after retreiving them from the NewsAPI database.

3.2.1 Code Snippet

```python
def get_nba_news(api_key, query="NBA", file_path="data.json"):
    url = 'https://newsapi.org/v2/everything'
    params = {
        'q': query,
        'language': 'en',
        'sortBy': 'relevancy',
        'apiKey': api_key
    }
    response = requests.get(url, params=params)
    if response.status_code == 200:
        data = response.json()
        articles = data['articles']
```

3.3 JSON

JSON (JavaScript Object Notation) is a widely used data format designed for lightweight data exchange between systems. It is text-based, human-readable, and easy for machines to parse, making it an ideal format for APIs and data storage. JSON is structured as key-value pairs which is similar to a Python dictionary, which makes it easy to work with in various programming languages. One of the main reasons JSON is so popular is its simplicity compared to other data formats like XML, which requires more complex parsing and has a heavier syntax. JSON is also language-agnostic, which means it can be used easily in a bunch of different programming environments, from JavaScript and Python to Java and C++. Because of its lightweight features, JSON is highly efficient in terms of data transmission. This efficiency is important when working with APIs like NewsAPI, where large amounts of data need to be retrieved quickly and processed efficiently. After retrieving my data from NewsAPI, I am saving my data in a json file. Below is an example of what the articles look like after being saved. When getting my sentiment analysis score, I use the content that is provided with each article, and I perform the sentiment analysis on that.

```
{
            "source": {
                "id": "espn",
                "name": "ESPN"
            },
            "author": "David Purdum",
            "title": "Vegas man arrested over
            Porter betting scheme",
            "description": "Shane Hennen was
            arrested and charged in
            connection to an illegal betting scheme
            involving former NBA player Jontay Porter.",
            "url": "https://www.espn.com/nba/story/_/id/43397591/
            // vegas-man-arrested-betting
            // -scheme-involving-jontay-porter",
            "urlToImage": "https://a1.espncdn.com/
            // combiner/i?img=
            // %2Fphoto%2F2024%2F0325%
            // 2Fr1309956_1296x729_16%2D9.jpg",
            "publishedAt": "2025-01-13T17:37:57Z",
            "content": "A Las Vegas man was charged in connection
            to an illegal betting scheme involving Jontay Porter,
            the former Toronto Raptors player
             for gambling purposes an\u2026 [+2522 chars]"
        },
```

3.4 Sentiment Analysis

When analyzing sentiment, I process the article content using spaCy's NLP pipeline, which tokenizes the text and prepares it for sentiment evaluation with SpacyTextBlob. SpaCy's NLP pipeline is a sequence of processing steps that transforms raw text into structured data for natural language understanding [2]. The pipeline typically includes tokenization, part-of-speech (POS) tagging, dependency parsing, named entity recognition (NER), and optional components such as lemmatization or custom-trained models. Tokenization, the first step in the pipeline, breaks the text into individual units called tokens, which can be words, punctuation marks, or special characters [2]. Unlike simple whitespace-based splitting, spaCy's tokenizer uses sophisticated linguistic rules and statistical models to handle contractions, multi-word expressions, and special cases like email addresses or dates. This structured tokenized output serves as the foundation for downstream NLP tasks, including sentiment analysis using tools like SpacyTextBlob, which assigns sentiment scores based on the processed tokens. After this process is conducted, only the polarity score is then extracted and calculated for teams mentioned in the articles. To accurately count team mentions, I iterate through each article's content and check if a team name appears, to make sure that occurrences are counted correctly. By preprocessing the data in this structured method, I improve the accuracy of both the mention counting and sentiment analysis processes by lowering the amount of errors and ensuring that the results are consistent. After collecting all of my data and processing it, I perform sentiment analysis and mention counters for each nba team. For the mention counter, the method is as simple as iterating through the articles and if a team is mentioned, I add 1 to that team's counter.

To analyze the sentiment of sports articles, this study utilizes two natural language processing (NLP) libraries: spaCy and TextBlob. spaCy is a strong NLP library created for text processing, while TextBlob provides a simple API for performing various text analytics, including sentiment analysis. In particular, TextBlob is used to calculate the polarity score of text blocks, which measures sentiment on a scale from -1 (highly negative) to 1 (highly positive), with 0 indicating a neutral sentiment[15]. TextBlob's sentiment analysis is powered by the Natural Language Toolkit (NLTK) and uses a pre-trained lexicon-based approach, where words are assigned predefined sentiment values based on their polarity in a sentiment lexicon. The polarity of a given block of text is then calculated by averaging the sentiment scores of individual words, considering their contextual weight and intensity modifiers. When looking at a word, it considers factors such as whether the word has increasing intensity (e.g., "very good" vs. "good") or whether it is negated by surrounding words (e.g., "not bad" vs. "bad"). Additionally, TextBlob assigns a subjectivity score ranging from 0 (objective) to 1 (highly subjective), which indicates the degree to which a text expresses opinion rather than factual information. The sentiment analysis method in TextBlob relies on the Pattern library's lexicon, which is based on a combination of machine learning and rule-based heuristics to determine sentiment values [15]. Below is an example of the flow of how Textblob works, and how it produces the polarity score and the subjectivity score. Textblob starts by taking in a piece of text of any size: in this case, it is a smaller piece of text. The next step is the tokenization step where you can see how Textblob breaks up the text as a whole and creates tokens of the individual characters. The next step is the POS or part of speech tagging where the algorithm will tag each token with a grammatical category. After that, each token is then given a polarity score based on a predefined sentiment lexicon. Lastly, as you can see in the diagram/flow chart below, the sentiment is calculated and the input text is finally given a polarity score and a subjectivity score. In this particular case, for this study, the only thing that is needed is the polarity score so that is what is extracted from the Textblob model and used for analysis. Below is a visual representation of the workflow and steps that Textblob goes through to get a polarity score from a section of text.

3.4.1 Flow Diagram of Textblob

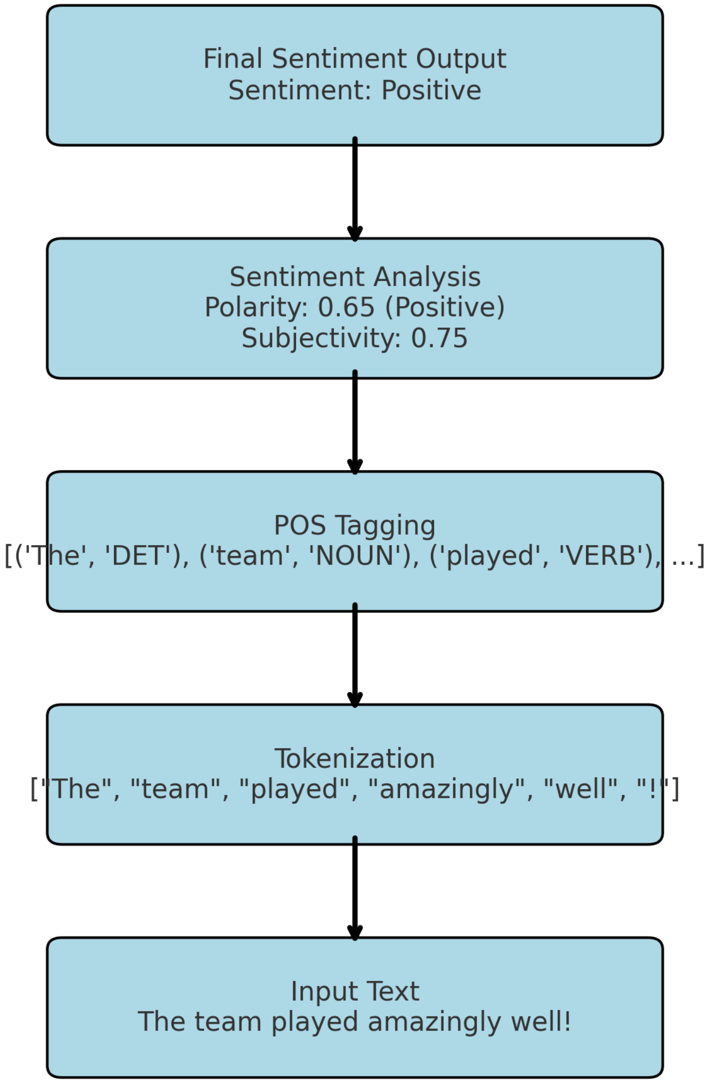
Final Sentiment Output
Sentiment: Positive
Sentiment Analysis
Polarity: 0.65 (Positive)
Subjectivity: 0.75
POS Tagging
[('The', 'DET'), ('team', 'NOUN'), ('played', 'VERB'), ...]
Tokenization
["The", "team", "played", "amazingly", "well", "!"]
Input Text
The team played amazingly well!

Figure 1: Textblob Pipeline

3.5 Design

My final product is a dashboard that displays graphs with each team's sentiment score and mentions count. To create the dashboard, I used a tool called 11ty to create a dashboard template with a few commands[13]. The dashboard template is created using the HTML(Hyper Text Markup Language) language. For the styling of my dashboard I created a CSS(Cascading Style Sheets) file that is embedded within the same folder that 11ty created when making the dashboard. Using 11ty was a very easy way to create a simple yet effective dashboard framework with all of the information on it that I needed, and all I had to do was add the CSS file myself.

For my dashboard, I decided to have the title of my project on the top of the dashboard and the two graphs that I am generating side by side right in the middle.This layout was chosen to ensure that the results are immediately visible to users, which will minimize the need for additional navigation when directed to the dashboard. I decided to go with a white background for the graphs and a black background for the dashboard with bright colors for the plots on the graphs. The selected color scheme enhances data visibility by ensuring that key results are immediately distinguishable from everything else and it is easy to understand what you are looking at. For the graph that shows each team and the amount of mentions that they got, I decided to go with a line chart with the number of mentions on the y axis and the NBA teams on the x axis. A line chart effectively visualizes the frequency of mentions for each team, facilitating comparison across all teams. For the graph that shows each team's sentiment scores, I decided to go with a bar chart that has each NBA team on the x axis and the range of sentiment scores on the y axis. A bar chart was chosen to effectively represent sentiment scores, as it clearly differentiates between positive and negative values by displaying bars above and below the 0.0 mark. This was an effective way to show which teams had a negative sentiment score and which teams had a positive sentiment score.

For this project, I utilized Plotly to create interactive visualizations that effectively present my data. Plotly is a powerful Python visualization library that allows for the generation of high-quality, interactive plots, making it ideal for my web-based dashboard. The key steps in constructing this graph were extracting NBA teams and their respective sentiment scores, using go.Bar() to generate a bar trace where the x-axis represents the team names and the y-axis represents their average sentiment scores, and customizing the chart with the correct labels, titles, and an angle adjustment for x-axis labels to make sure that the graphs are readable. This resulted in a bar chart that allows users to quickly compare how different teams are perceived based on sentiment analysis. To visualize how frequently each NBA team was mentioned in the dataset of articles, I created a line chart using go.Scatter(). This chart follows a similar data extraction process: extracting the mention counts for each team, using go.Scatter() to generate a line plot incorporating markers for better visibility, and adjusting the layout to include the correct titles and label arrangements. The line chart allows users to track which teams are most frequently mentioned, helping to identify trends in NBA media coverage. One of the major benefits of using Plotly is its ability to export plots as JSON using plotly.io.to_json(). This feature allowed me to save my plots as JSON files, which I could then dynamically load into my web-based dashboard using JavaScript. The process involved converting each Plotly figure to a JSON string, saving these JSON objects to a file (plot_data.json), using JavaScript's fetch() function to retrieve and parse the JSON data in an HTML file, and rendering the graphs in the browser using Plotly.newPlot(). This method made sure that my visualizations remain interactive and can be updated dynamically without requiring the user to re-run the Python script. Plotly offers several advantages over traditional static plotting libraries. The interactive nature of the graphs allows users to hover over data points, zoom in/out, and explore data trends more effectively. Additionally, its compatibility with both Python and JavaScript enables me to easily integrate it with web applications, making it an excellent choice for my project. Below is an example of what the graphs look like on the dashboard.

3.5.1 Graph Designs

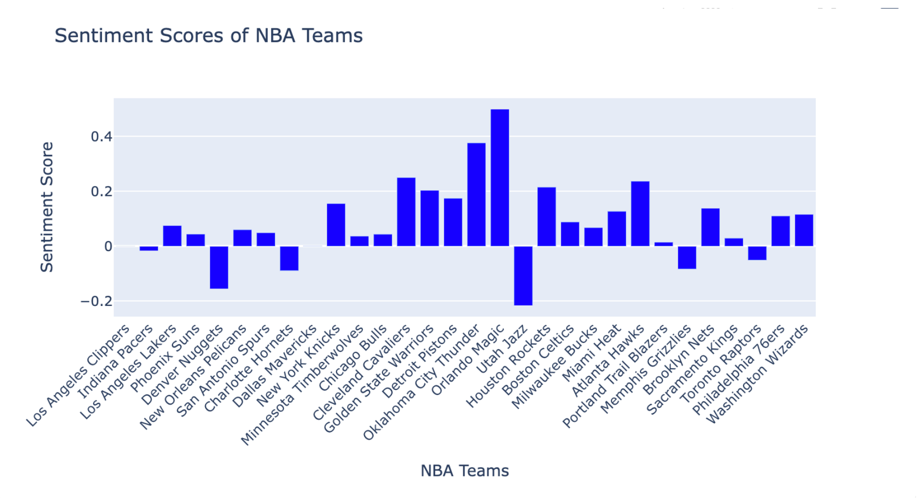

Figure 2: Sentiment Score Dashboard Graph

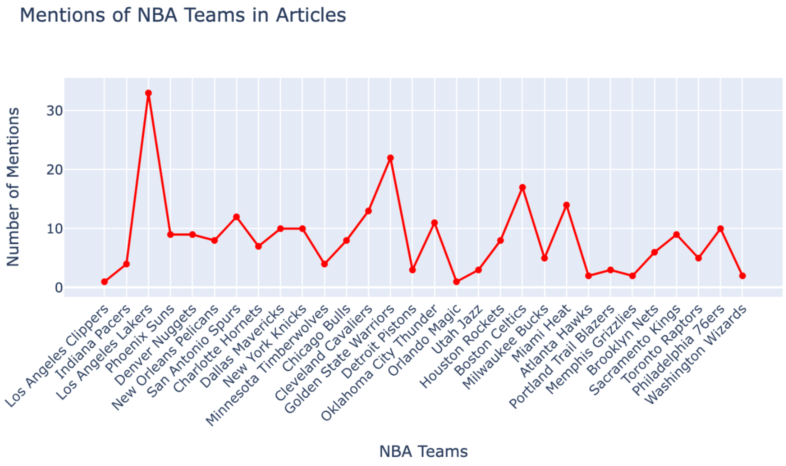

Figure 3: Mention Dashboard Graph

4 Experiments

4.1 Data Manipulation Experiment

For the first experiment in my project, I to explored how the volume of data, in this case, the number of NBA articles analyzed, impacts the accuracy and outcomes of sentiment analysis and mention count. The overall goal of my project is to perform sentiment analysis and a mention count on NBA articles using a large dataset gathered from NewsAPI, which is a service that provides access to various news sources and articles. In this experiment, I manipulated the size of the dataset by running sentiment analysis on a number of distinct groups of articles: some larger datasets of articles, and some smaller datasets of articles. By comparing the results of these two sets, I determined whether there are noticeable differences in sentiment scores and mention counts, as well as how the volume of articles affects the reliability and consistency of the results.

To start, I collected a batch of articles from NewsAPI, ensuring that they are NBA-related and come from a variety of different sources to provide a broad spectrum of viewpoints. The selection spanned multiple months and cover a range of topics, such as match results, player interviews, and NBA drama, to ensure the dataset reflects the diversity of the NBA. The larger datasets consisted of a significant number of articles, typically in the thousands, to represent a more generalized and robust view of the sentiment within the NBA media. The smaller datasets, on the other hand, only include da select number of articles, likely in the range of 100 to 500. The reason for this is to simulate how sentiment analysis performs with a more limited amount of data while still reflecting the sentiment of the articles.

The experiment involved running sentiment analysis on all of the sizes of datasets using the same sentiment analysis model and methodology. I employed a pre-trained natural language processing (NLP) model that is designed for sentiment analysis called TextBlob which is well-suited for analyzing text from social media or news articles about anything. The models work by analyzing the language used in the articles and assigning a sentiment score that ranges from negative to positive, with neutral sentiment falling in the middle. Once I have applied sentiment analysis to the articles, I evaluated the results by examining the distribution of sentiment scores and mention counts for all of the different sizes of datasets.

By comparing the sentiment scores of the larger and smaller datasets, I hope to uncover any trends or significant differences. For example, it is possible that the sentiment scores derived from the large dataset was be more balanced, reflecting a more fair and representative view of sentiment across the NBA media landscape. On the other hand, the smaller dataset may yield more extreme results, either overly positive or negative, simply due to the limited sample size. I anticipate that the large dataset may also produce more stable and reliable sentiment scores, as the variability introduced by individual articles in a small dataset could have a larger impact on the final results. Also, with a larger dataset, there is obviously be a noticeable difference in the amount of mentions each team received. This was be another thing to look at and see if the amount of mentions for the teams has an effect on the sentiment scores.

Additionally, I plan to examine how the difference in sentiment scores changes with the dataset size. For example, in the small dataset, there could be outliers that vary the overall sentiment, whereas in the larger dataset, these outliers may be averaged out, leading to more consistent findings. This comparison helped me understand whether sentiment analysis on a small set of articles is prone to producing more skewed results, which could potentially influence how the analysis is interpreted in different contexts. Another important thing to think about in this experiment is the speed and efficiency of running sentiment analysis on different sizes of datasets. The large dataset required more processing power and time to analyze, potentially affecting the performance of the sentiment analysis model. By looking at how the time and resources required to run sentiment analysis scale with the dataset size, I can determine whether or not there are benefits to using a larger dataset versus a smaller dataset.

Finally, the results of this experiment provided valuable information into the trade-offs between dataset size and sentiment analysis accuracy. It allowed me to determine whether running sentiment analysis on a large amount of data showed more reliable, consistent, and meaningful results compared to with smaller datasets. The experiment also showed how the computational costs associated with analyzing large datasets versus smaller datasets influenced the ability to perform sentiment analysis in real-world applications.

4.1.1 Data Manipulation Results

This experiment's main goal is to compare sentiment scores and mention counts of NBA teams by running sentiment analysis on a smaller dataset versus a larger dataset. The findings from running the sentiment analysis on both datasets reveals keys into how data size impacts sentiment results, the frequency of mentions, and the overall reliability of conclusions drawn from running sentiment analysis on different volumes of data. The first dataset had significantly fewer articles than the second, which directly influenced both sentiment scores and mention counts. Some teams were completely absent from the first dataset, receiving no mentions, while their presence in the second dataset resulted in measurable sentiment scores. Additionally, teams that appeared in both datasets showed shifts in their sentiment score, which shows how larger datasets help clear up sentiment analysis results.

One of the largest impacts of dataset size was on sentiment scores. In the smaller dataset, sentiment scores varied widely, with some teams having extremely high or low scores making them outliers. For instance, the Houston Rockets had a very high sentiment score of 0.8000, which is abnormally high in real-world sentiment scenarios. On the other hand, the Denver Nuggets had a negative sentiment score of -0.2189, suggesting strongly negative coverage. However, in the larger dataset, the Houston Rockets' score dropped to 0.1321, and the Denver Nuggets' score improved to -0.1378. This demonstrates that a larger dataset reduces the impact of outliers and extreme sentiment values, leading to more balanced and reliable results.

Another major issue with using a small dataset is that some teams were completely missing from the analysis. The Boston Celtics, Chicago Bulls, and Brooklyn Nets, for example, had no mentions in the small dataset, making it impossible to analyze their sentiment, and in the mentions chart there were no values for those teams. However, in the larger dataset, these teams had measurable sentiment scores, with the Boston Celtics receiving a score of 0.0671, the Chicago Bulls at 0.0588, and the Brooklyn Nets at 0.0867. This finding emphasizes the risk of drawing conclusions from small datasets, because teams with fewer media mentions may not be represented at all, leading to an incomplete graph with misleading information. The larger dataset ensured that all teams received some level of coverage, making the results more representative of overall media sentiment in the NBA.

In addition to missing teams, sentiment scores were changed in teams that were present in both datasets. Some teams saw their sentiment scores shift significantly as more data was included. The Dallas Mavericks had a sentiment score of -0.0316 in the small dataset, which slightly improved to -0.0237 in the large dataset. The Oklahoma City Thunder's sentiment score increased from 0.3531 to 0.4187, while the Phoenix Suns had one of the most notable changes, going from a negative sentiment score of -0.1500 in the small dataset to a positive score of 0.0687 in the larger dataset. These shifts show that sentiment trends can be misleading when sentiment analysis is run using a small dataset.

The size of the dataset also significantly impacted mention counts. Predictably, a larger dataset led to higher mention counts even though it was distributed differently across the teams. Teams with lower mentions in the small dataset gained more representation in the larger dataset. For example, the Los Angeles Lakers had eight mentions in the small dataset but jumped to 30 mentions in the larger dataset. Similarly, the Golden State Warriors increased from five mentions to 18, while the Boston Celtics, who had zero mentions in the small dataset, had 11 mentions in the larger one. This increase highlights how smaller datasets may underrepresent certain teams, which will give results that aren't representative in the analysis. A larger dataset provides a more inclusive view of media coverage, making sure that all teams are fairly represented.

For teams with very low mention counts in the small dataset, a larger dataset made the difference between a team having no data points to perform sentiment analysis. The Detroit Pistons and Toronto Raptors, for example, had zero mentions in the small dataset but gained two and five mentions, respectively, in the larger dataset. The Brooklyn Nets also went from having no mentions to five mentions. This shows that using a small dataset can lead to a team having no data points which can make people think that certain teams are not being covered at all when they are only being underrepresented because of the small amount of articles that the sentiment analysis is being performed on.

On the other hand, some teams had high mention counts in both datasets, but they still improved from increased size of data. Popular teams like the Los Angeles Lakers, Miami Heat, and Golden State Warriors received a significant amount of media coverage regardless of dataset size. The Lakers increased from eight mentions in the small dataset to 30 in the large one, while the Miami Heat jumped from three mentions to 12. The Golden State Warriors had five mentions in the small dataset but 18 in the larger dataset. This suggests that while top teams are likely to be covered no matter the size of the dataset, the number of mentions they receive still increases with a larger dataset, which allows for better results when performing my analysis.

The findings from this experiment really emphasize the importance of using large datasets when performing sentiment analysis. A small dataset can lead to confusing sentiment results due to extreme outlier sentiment scores, missing teams, and skewed mention counts. A larger dataset will make sure the results are more accurate. For example, the Houston Rockets' weirdly high sentiment score in the small dataset was balanced out when more articles were added, reducing the risk of over exaggerated information based on limited data. Similarly, teams that were absent from the small dataset were accounted for in the larger one, which makes it a more inclusive analysis.

To make sure I have the best results when performing sentiment analysis moving forward, it is important to use the largest dataset available. Larger datasets capture a broader range of articles, reducing the influence of individual outliers and providing a more balanced view of media sentiment in the NBA. Monitoring how sentiment scores shift as dataset size increases help me realize trends that showed me larger datasets are more effective. By using more articles and a larger dataset, sentiment analysis can provide a more data-driven and reliable view of how NBA teams are perceived in media coverage.

4.1.2 Data Manipulation Graphs

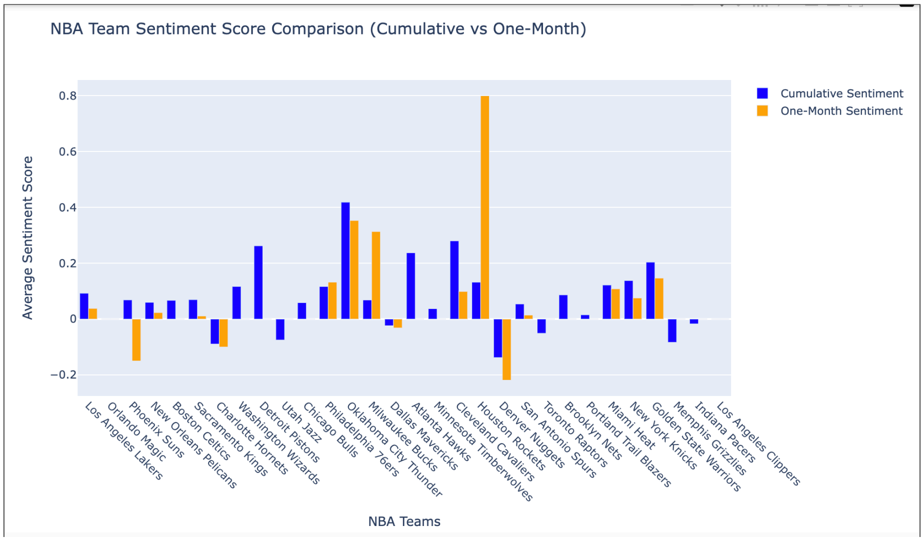

Figure 4: Mentions

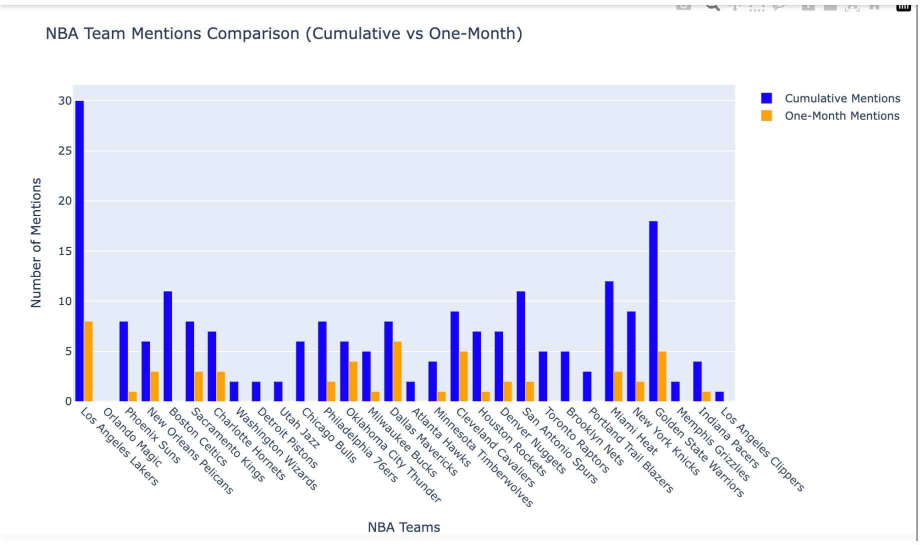

Figure 5: Sentiment Scores

4.2 Sentiment Analysis Tool Experiment

For my second experiment, I evaluated two different sentiment analysis tools, SpaCyTextBlob and VADER, to determine which one performs better in analyzing the sentiment of NBA-related sports articles. This experiment compared their accuracy, reliability, and overall performance in their abilities to collect the sentiment scores in the most efficient way possible.. Since the articles are sourced from a bunch of different media outlets over a five-month period, this experiment told me how each tool handles sports-related tone and language, which consisted of things like sports phrases, different emotional tones, and things like idioms.

SpaCyTextBlob is a sentiment analysis tool that integrates TextBlob's sentiment capabilities directly into the spaCy NLP pipeline. It uses a lexicon-based approach, which means that is gives polarity scores to words and averages those scores across the text, in this cas, NBA articles. SpaCyTextBlob is known for how simple it is and how it can handle structured language effectively, making it a strong choice for analyzing well-written articles such as NBA news pieces. However, SpaCyTextBlob may struggle with detecting nuanced emotions, sarcasm, or slang often present in sports media. On the other hand, VADER (Valence Aware Dictionary and sEntiment Reasoner) is another lexicon-based sentiment analysis tool designed specifically for social media text. While VADER is part of the NLTK library, it is widely used across domains due to its ability to capture both positive and negative sentiment while understanding context. VADER assigns intensity scores to individual words and combines them to produce an overall sentiment score. VADER is very good at detecting casual language, short phrases, and highly emotional content which are traits that are common in sports reporting.

To conduct this experiment, I am going to use the same dataset of NBA-related sports articles using both sentiment tools. The dataset includes five months' worth of articles stored in a JSON file called cumulativedata.json. Each article contains content that was processed by both SpaCyTextBlob and VADER to produce sentiment scores. By running both tools on identical content, I can directly compare their outputs and results which helped me identify patterns in the sentiment scores. By keeping all other variables the same including the data, this allowed me to directly see how different sentiment analysis tools produced different sentiment scores.

After preprocessing the data, I continued with performing sentiment analysis. I used my existing sentiment analysis file, which uses and includes the SpaCyTextBlob tool, to generate polarity scores for each article. These scores were then be stored in a csv file that tracks the average polarity score for each NBA team based on the articles that were just analyzed. To compare results, I used my newly created sentiment analysis VADER file, which implements VADER to generate separate sentiment scores in a similar method. Just like with SpaCyTextBlob, the VADER results were also be separated by NBA team which allows for direct comparison. After collecting both sets of results, I visualized the findings to highlight key differences. I created two bar charts, one displaying the sentiment results from SpaCyTextBlob and the other showing VADER's output. This visual comparison made it easier to identify differences between the different teams sentiment scores, such as whether one tool consistently finds more extreme, neutral, or balanced scores. Additionally, I examined whether one tool produces scores that better align with expected trends based on real-world NBA events, such as winning streaks, trades, or controversies. The overall goal of the project is to try to prove media bias in the NBA so comparing the actual results to the trends in the NBA helped me figure out which tool is better. To analyze the results, I compared the ranges of sentiment values produced by each tool for each NBA team. If one tool consistently assigns really high or really low sentiment scores, this may show a systematic bias or a different sensitivity to the language including things such as sarcasm.

I can estimate that SpaCyTextBlob produced smoother, more balanced sentiment scores because of its focus on structured language. Since news articles are normally more formal and less casual than social media content, SpaCyTextBlob's lexicon may be better for this specific datset. However, I expect VADER to be better at identifying stronger emotional language, particularly in bold headlines, opinion pieces, or things such as sarcasm within the articles. Depending on the results, I may choose to replace SpaCyTextBlob with VADER if the results show me that VADER is a more effective tool for the tasks that I am completing. This experiment gave a deeper understanding of how different sentiment analysis tools interpret sports-related text, specifically articles about the NBA. By directly comparing the performance of SpaCyTextBlob and VADER, I look to find the most effective tool for my project's dataset and improve the accuracy of my NBA sentiment analysis dashboard. The results of this experiment gave me the information I need to continue with my methods comfortably knowing that I am using the correct tools.

4.2.1 Sentiment Analysis Tool Results

The results of my experiment comparing SpaCyTextBlob and VADER for sentiment analysis in NBA-related sports articles revealed that SpaCyTextBlob is the best option for performing sentiment analysis for my project. The main goal of this experiment was to figure out which tool better captured the sentiment that is in the articles, particularly given the writing style and context that is found in sports reporting. After analyzing the data, SpaCyTextBlob consistently produced more balanced and reliable sentiment scores compared to VADER. VADER had some strengths when dealing with emotions and sarcasm, but SpaCyTextBlob's overall performance worked better with the NBA articles and data that I collected from NewsAPI.

One of the key and most identifiable differences between the two tools was the consistency in the polarity score or sentiment score. SpaCyTextBlob's polarity scores were relatively normal and even, even for teams that had a good amount of media coverage. For example, the New York Knicks had a SpaCyTextBlob average polarity score of 0.1379, while VADER assigned them a higher score of 0.2383. Similarly, the Los Angeles Lakers, a team frequently in the spotlight, had a score of 0.0924 with SpaCyTextBlob, compared to 0.2721 with VADER. These differences show that VADER may exaggerate positive sentiment, possibly from looking at the articles and analyzing things like sarcasm and giving the text a very positive score because of that. In sports articles, there is a lot of excitement, speculation, and dramatic language which can create outliers and weird sentiment scores if not carefully managed. In contrast, SpaCyTextBlob's more conservative scoring seemed to better reflect the tone of these NBA articles.

Another important thing to look at was that VADER had a wider range of scores, with way larger positive and negative values. For example, the Boston Celtics received a high VADER score of 0.4764, significantly higher than their SpaCyTextBlob score of 0.0671. These results can tell us that VADER was more sensitive to emotional language such as opinions, which may not always give off the intended sentiment of the article. Sports articles often emphasize dramatic language to get readers to read the articles, and VADER's tool has the tendency to not take into account these tones which leads to extreme sentiment scores whether that be negative or positive. SpaCyTextBlob's more balanced methods avoided these exaggerated scores and better showed the neutral tone that was actually in these articles. Negative sentiment scoring was another area where SpaCyTextBlob outperformed VADER. Teams like the Utah Jazz and Toronto Raptors received negative scores from both tools, but SpaCyTextBlob's results better aligned with what was expected of the scores. The Utah Jazz had a SpaCyTextBlob score of -0.0750, while VADER gave them a more extreme score of -0.2202. Similarly, the Toronto Raptors had a score of -0.0509 with SpaCyTextBlob compared to -0.1235 with VADER. Both tools detected negative sentiment, but VADER's exaggerated negativity score was too extreme taking in the context of the articles and what they consisted of. In my own review of articles discussing these teams, the language was often critical but not very negative like the VADER score suggested. SpaCyTextBlob's moderate scoring better showed and displayed what the articles actually consist of.

Additionally, SpaCyTextBlob's results worked better with the number of article mentions, which further proved that it was the better tool. Teams like the Los Angeles Lakers, who had the highest mention count (30 mentions), received a polarity score of 0.0924 from SpaCyTextBlob which says they have a slight positive sentiment that reflects their strong media presence. VADER on the other hand assigned them a higher score of 0.2721 which seems high given that they were mentioned 30 times. Similarly, teams with fewer mentions, such as the Atlanta Hawks (2 mentions), received lower sentiment scores in both tools. After reviewing these articles, SpaCyTextBlob's sentiment scores aligned better with how these teams were portrayed in the articles that they were mentioned. This stronger alignment with mention frequency suggests that SpaCyTextBlob's scoring was better given the context of how many times each team was mentioned than VADER.

While VADER demonstrated some strengths, particularly in capturing emotional language such as sarcasm, it proved that it was less fitting for my project. Sports journalism often relies on excitement and dramatic language to get readers attention, and VADER's sensitivity to these articles led to frequent scores that were either way too high or way too low for what was said in the articles. Also, SpaCyTextBlob's more measured way of collecting sentiment provided results that better reflected the true sentiment in the articles. For my NBA sentiment analysis dashboard, accuracy and consistency are important to ensuring the visualized data is meaningful and reliable. SpaCyTextBlob's more moderate, controlled scoring methods makes sure that the resulting sentiment patterns are less messed up by exaggerated emotions or misclassified content.

In conclusion, this experiment clearly showed that SpaCyTextBlob is a better fit for my project than VADER. Its balanced polarity scores, improved handling of neutral language, and closer alignment with mention frequency produced more accurate and meaningful sentiment results are all reasons why SpaCyTextBlob is better for my project. While VADER may excel in social media analysis or highly emotional content, SpaCyTextBlob's strengths in handling structured language and sentiment make it the best tool for analyzing NBA-related sports articles. Moving forward, I will continue using SpaCyTextBlob as the primary sentiment analysis tool in my project to ensure my dashboard provides clear and accurate insights into how NBA teams are portrayed in the media.

4.2.2 Sentiment Analysis Tools Results Graphs

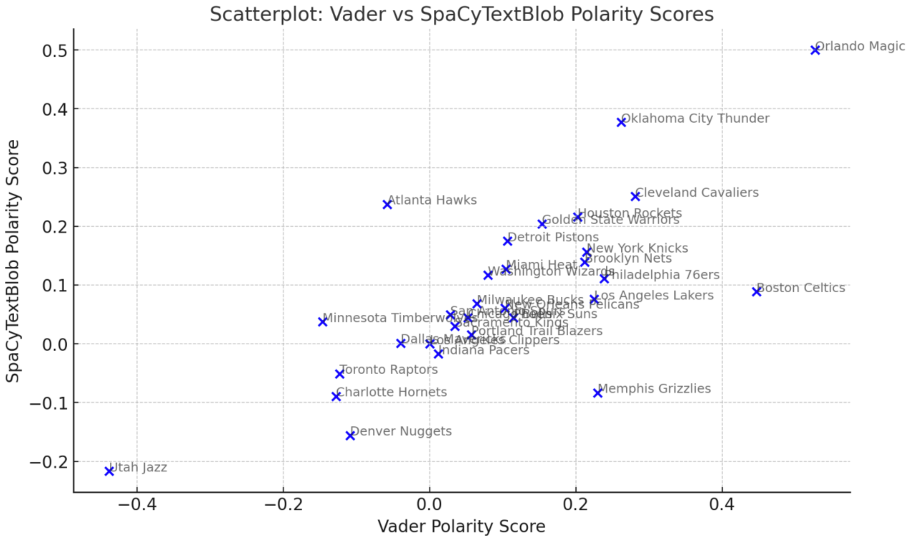

Figure 6: Different Sentiment Analysis Tools

4.3 Threats to Validity

There are a couple things that could threaten the validity of the conclusions drawn from both experiments. In the first experiment, which compares small and large datasets, a key issue is sampling bias. The articles sourced from NewsAPI may not be fully representative of all NBA-related media coverage, as certain sources or topics could be overrepresented. This could lead to biased sentiment scores if, for example, popular media outlets with specific tones are the only outlets that are included in the datasets. Additionally, for the sentiment tool experiment, tool limitations are a large threat to validity. Both SpaCyTextBlob and VADER have constraints in processing sports articles that may lead to the experiment not actually testing how they perform on sports articles. For example, SpaCyTextBlob, may struggle with sarcasm or emotional exaggeration which is very prevalent in sports articles. VADER, on the other hand, can sometimes overemphasize emotional language, leading to extreme sentiment scores. These limitations could introduce inconsistencies between the two tools, skewing the experimental results.

Another threat to validity is text variability. NBA articles differ greatly in the content and tone. Things such as player performance reports to trade rumors and team controversies can all be topics of NBA related articles. These variations can affect how sentiment analysis tools interpret the text, leading to inconsistent sentiment scores across articles of different types. The way in which the data is preprocessed for each tool also introduces potential biases. Differences in tokenization, stop-word removal, or other preprocessing steps could lead to disparities in how sentiment is measured across the two tools, further complicating the results. Outliers are another threat in both experiments. Both small and large datasets can include articles that use sensational language or exaggerated tones, which could distort the sentiment analysis, especially in smaller datasets where outliers have more influence. In the larger dataset, outliers may have less impact, but their presence is still a concern when interpreting results. Additionally, since the articles span across several months, temporal changes in NBA events, such as player injuries, trades, or key team victories, may affect sentiment analysis in ways that are difficult to control for which could potentially introduce bias.

The second experiment, which compares SpaCyTextBlob with VADER introduces specific risks related to the tool comparison. The results from both tools may not be directly able to be compared because their approaches to sentiment analysis are different. VADER is designed to capture more extreme sentiment expressions, particularly for social media-like content, while SpaCyTextBlob may struggle to pick up on those same types of expressions. This could result in differences in sentiment scores that are not necessarily related to the dataset itself but are just differences in the tools and the way that they calculate sentiment analysis.

5 Future Work

5.1 Summary of Research

My project Hoop Hype is designed to evaluate the sentiment and media coverage of NBA teams based on articles retrieved from NewsAPI. The main part of the project involved three primary steps. The three steps are data collection, sentiment analysis, and data visualization. Each of these steps used specific technologies and methods to make sure the data is accurately represented. I started my data collection by retrieving articles from NewsAPI using a specific query, which allowed targeted searches for NBA-related content. By setting the query to "NBA," the system filtered results to only include articles that directly referenced the NBA, which will guarantee that the collected data was highly relevant to the project's goals. The data was retrieved in JSON format, which is a good structure that is the best option for text-based data processing. JSON allows for the storage of large amounts of data, content, and source information, which makes it possible to iterate through and analyze large volumes of data efficiently. Once the articles were collected, the data was structured to make sure that each article's content could be easily processed for sentiment analysis.

The sentiment analysis step used SpaCy's NLP pipeline enhanced with SpaCyTextBlob to perform the sentiment analysis. The combination of SpaCy's pipeline and SpaCyTextBlob provided a strong text-processing solution which performed tokenization, lemmatization, and sentiment scoring. Sentiment scores were calculated using polarity values, which ranged from -1 to +1. Each article's content was processed, and NBA teams mentioned in the article were cross checking using a CSV dataset that contained a complete list of NBA teams and players. This cross checking step ensured that sentiment values were attributed to the correct teams and that each team was accounted for. For each identified mention, the polarity score was calculated and averaged to provide an overall measure of how each team was portrayed in media coverage. The sentiment analysis step provided important insights, which revealed very large differences in sentiment between teams. For instance, the Oklahoma City Thunder had a remarkably high positive sentiment score, while the Denver Nuggets had a significantly negative sentiment score. These patterns highlighted the different media narratives for each NBA team.

To make sure that my results were displayed correctly, the data was presented using two bar charts, one displaying the average polarity score for each NBA team and another showing the number of mentions each team received. This visualization method allowed users to easily compare team sentiment and team mentions within media coverage. By presenting the data in a clear and accessible format, Hoop-Hype facilitated a better understanding of how media coverage might influence public perception of NBA teams incorrectly. The bar charts also highlighted the differences in team mentions, showing that high-profile teams like the Los Angeles Lakers had greater media coverage than smaller market teams. This clear visual representation allowed users to quickly identify key findings and patterns in the data. Future Areas of Work

5.2 Future Areas of Work

While the Hoop Hype tool showed valuable information and findings, there are several opportunities for future improvement and more research to be done. One direction that someone continuing this research could go in is enhancing the sentiment analysis model itself. Although SpaCyTextBlob effectively analyzed sentiment, someone could either incorporate multiple sentiment analysis tools or create their own tool for calculating sentiment analysis. Using multiple sentiment analysis tools could allow for more accurate results by combining the scores of each tool.

Another potential enhancement is the inclusion of calculating sentiment analysis for individual players in the NBA. While Hoop Hype currently evaluates sentiment at the team level, analyzing individual players would offer more information about media bias and individual narratives. Player level sentiment data could reveal which athletes are portrayed positively or negatively, providing valuable information about which players are liked or disliked for people such as fans, analysts, and journalists. Expanding the tool to evaluate individual mentions would be very similar to my method for the teams, except you are cross checking the articles with a CSV file that consists of NBA players instead of NBA teams.

Additionally, introducing analysis of sentiment over time would provide new findings into how media sentiment shifts over time in the NBA. Tracking sentiment analysis patterns across multiple months or seasons could identify change in sentiment scores which could be linked to team performance, trades, or controversies. For example, a positive sentiment score might align with a winning streak, while a sentiment score decline could indicate a major team scandal or injury report. Implementing a time series visualization within Hoop-Hype's dashboard would allow users to explore these changes dynamically which adds an important dimension to the analysis. Without knowing why a team's sentiment score spiked really high or fell really low, it might seem like the media is just being biased for no reason, and something like this would identify the real media bias in the NBA by allowing users to see when sentiment scores were calculated.

Expanding Hoop Hype's coverage to include multilingual articles about the NBA is another promising way for future work to be done. Since the NBA has a global fan base, analyzing articles written in other languages could capture media narratives from international sources. This would provide a more overall view of global perceptions surrounding NBA teams. Integrating multilingual support would require additional language models and translation tools to ensure accurate sentiment scoring across non-English content. There are NBA games that are played in other countries, and there is a lot of media coverage surrounding those games and players. Basketball, especially the NBA, is a global sport and including articles from other countries and in other languages would be a very good extension to Hoop Hype.

Another thing that could be added to this project in the future would be the improvement of the visualization aspect of the project. Based on what my dashboard already looks like, I could add more visuals/graphs to display even more information about the NBA teams and their sentiment scores. For example, trend lines or average sentiment benchmarks would be very helpful to users when they are navigating the dashboard. Going off of that, the visualization could also include more interactivity. This would allow for the users to learn more about what they want to instead of not having many options when it comes to what they are looking at.

Finally, expanding the dataset by adding data from social media platforms such as Twitter, Reddit, or Instagram would provide valuable insights into fan sentiment and discussions. Social media often reflects real time reactions to NBA events, trades, or performances which makes it an important data source for understanding public perception. A lot of the NBA media on large television networks have social media accounts, so this would be a very good addition to this project. The integration of social media content into Hoop Hype would involve implementing API connections to platforms like Twitter, parsing things such as emojis and hashtags, and extracting user sentiment for improved analysis.

5.3 Ethical Issues

While Hoop Hype offers valuable insights into NBA media coverage, the project raises several unresolved ethical issues that this future works section is going to address. One notable issue is the potential bias in article selection. Since NewsAPI's filtering methods prioritize relevance and popularity, certain viewpoints may be overrepresented, while others may be excluded. This creates a risk that extreme narratives or biased reporting may incorrectly influence sentiment scores. For example, if major publications have a negative stance toward a specific team, this bias could skew Hoop Hype's sentiment analysis results. The main sources being used for this project are from major sports news sources such as ESPN and Bleacher Report. By not including smaller and not as popular new sources, the results could be a direct reflection of the bias of the new sources from NewsAPI.

Additionally, automated sentiment analysis models like SpaCyTextBlob face limitations when processing complex language patterns. Sports journalism frequently includes things such as sarcasm, hyperbole, and metaphor which are styles that can confuse NLP models. Even after testing a different NLP model and choosing the best one, it still isn't always accurate. For instance, a sarcastic remark about a player's performance may be interpreted as negative when the intent was for the comment to be funny. These limitations raise ethical concerns about the accuracy and reliability of sentiment analysis results. To fix this issue, future versions of Hoop-Hype and people who want to continue on with this work could employ tools that interpret the text before the sentiment analysis is calculated to provide context around sentiment scores.

The project also introduces potential data privacy concerns. While NewsAPI articles are publicly accessible, manipulating and analyzing this content at this large of a scale may raise questions about the news sources and the authors intent. News organizations may object to large-scale aggregation without proper attribution or consent. Future versions and continuations of the project Hoop Hype should consider methods to acknowledge and credit content creators appropriately while ensuring responsible use of the articles in sports media. Bias in sentiment models is another issue that raises concern. NLP models are trained on large text corpora, which may contain biases in language representation. Since NBA-related content may already reflect racial, cultural, or regional biases, these patterns could unintentionally influence Hoop Hype's sentiment scores. To reduce this risk, future versions of Hoop Hype should expand their training datasets by including their own training method which incorporates text that is similar to the text that is in sports media.

Finally, publicizing sentiment scores presents a risk of creating negative stereotypes or unfairly giving players, teams, or organizations bad reputations. A consistently negative sentiment score for a team or player could create biased narratives, even if the content lacks genuine criticism. To address this, future iterations of Hoop Hype could provide disclaimers that talk about the limitations of sentiment analysis tools and encourage the users to look at the data and interpret it on their own. Adding context alongside the sentiment scores may help users better understand the factors influencing sentiment results. By addressing these unresolved ethical concerns, Hoop Hype can improve its fairness, reliability, and social responsibility while continuing to provide meaningful findings about the NBA and their media coverage. One of the main purposes of this project is to try to prove media bias, but it needs to be done fairly and accurately without slandering NBA teams or players. There is a large ethical dilemma when it comes to the results of Hoop Hype exposing NBA teams and how they are portrayed in the media which will also expose the media by showing how they are reacting and writing about specific NBA teams.

6 Conclusion

The main purpose of this study is to set out to examine the presence of media bias within the National Basketball Association (NBA) by analyzing sentiment and mention counts of NBA teams and players in news articles. By using Natural Language Processing (NLP) techniques and sentiment analysis, the project went over different pieces of media coverage, showcasing differences in how teams and players are portrayed in the media. The data was visualized on a dashboard, allowing for a clear and unbiased representation of media sentiment and frequency of mentions. The results of the study display that media coverage within the NBA is not fair. Certain teams and players receive more attention and are covered in a predominantly positive or negative manner, which can raise questions about media bias allowing for professional or personal interpretations to be made. The findings align with concerns that media narratives significantly influence public perception and can impact teams and players beyond just their on-court performances. Additionally, the study uncovered that while sentiment analysis is a useful tool in identifying general trends in media coverage, it is not without its limitations. The inability of sentiment analysis tools to detect different text such as sarcasm or contextual implications means that some bias could still go undetected. Furthermore, the reliance on NewsAPI as the sole data source means that coverage was limited to specific outlets, potentially excluding other relevant media sources including things such as social media posts or small new outlets. Despite these limitations, the project successfully highlighted differences in media coverage which provided valuable insights into the ways in which media shapes public perception of NBA teams and players.

This study adds to the large discourse on media bias, particularly in the area of the NBA. While previous research has explored media influence in politics and entertainment, this study adds to the existing body of knowledge by demonstrating that similar patterns exist in sports journalism. The findings suggest that media outlets may favor certain teams or players, consciously or unconsciously, leading to unfair portrayals of these teams or players in the media for everyone to see. These findings support ideas of people in the media having an agenda and possibly framing these players or teams which emphasizes the role media plays in shaping public opinion.

From a more practical standpoint, this study provides valuable insights for NBA teams, players, and league executives. Understanding how media coverage varies across different teams can help organizations change their media strategies to make the negative portrayals better or leverage positive coverage to their advantage. Additionally, players and their representatives can use the findings to navigate public narratives more effectively by addressing unfair bias or capitalizing on positive sentiment. For journalists and media outlets, the study serves as a reminder of the responsibility they hold in ensuring fair and balanced reporting. Bias, whether intentional or unintentional, can have real-world consequences including things such as influencing public perception, player contracts, sponsorship deals, and even team success. The insights from this study can be used as a tool for self-reflection within the media industry which would encourage more efforts from these media outlets to consider things like bias if they want to maintain their integrity.

The NBA is a multi-billion-dollar industry with global influence, and the media plays a large role in shaping the landscape of the NBA. Media bias can impact revenue streams for teams which could affect merchandise sales, ticket sales, and sponsorships. If certain teams or players receive unfair negative coverage, it could scare away potential investors or brand partnerships. Conversely, teams that receive excessive positive coverage may gain an unfair advantage in marketability. From a social perspective, biased media narratives can influence fan behavior, leading to unfair criticism or support for players and teams based on skewed representations rather than what is actually going on with the specific players or organizations.

Despite the valuable findings that this study has given us, it is not without its limitations. One of the main limitations is the reliance on sentiment analysis tools, which are not always correct. These tools may misinterpret sarcasm, metaphors, and contextual subtleties, leading to occasional misclassifications of sentiment. While every effort was made to validate the accuracy of sentiment scoring, it is not guaranteed that the results are 100% accurate. Another limitation is the data source. NewsAPI was used to collect articles, but this does not include all possible media outlets covering the NBA. Some articles may not have been collected, and alternative sources, such as social media or independent blogs, were not included in the analysis. This means that the study presents only a partial view of media coverage and may not fully reflect the larger media landscape. Additionally, this study focuses exclusively on NBA teams and players, leaving out other professional basketball leagues or sports in general. While the findings are relevant within the NBA, they may not necessarily apply to other sports or leagues with different media dynamics. Future research could expand the scope of the project to include an analysis between different sports leagues to determine whether similar patterns exist which would validate the methods even more.

Given the limitations that were just mentioned, future research should consider several areas for further exploration. One path that could be taken is improving the sentiment analysis methodology by adding more advanced NLP techniques, such as higher level embeddings and deep learning models, to better capture sentiment scores and ensure the models accuracy. Additionally, expanding the dataset to include a wider range of news sources, as well as social media platforms, could provide a more comprehensive view of media coverage. Another potential area for future research is analyzing the correlation between media sentiment and different external factors such as team performance, market size, or player controversies. By analyzing and understanding the relationships between these things could allow you to have a deeper understanding about the media bias. Additionally, doing interviews with journalists and media professionals could provide information into the decision-making processes behind sports reporting and whether conscious biases exist or whether the bias is happening unconsciously. Finally, an important area for future research is examining the impact of media bias on fan engagement and perception. A survey-based study could measure how fans interpret and respond to media coverage, shedding light on whether biased reporting influences fan loyalty, player popularity, or even game attendance. Exploring these behaviors and analyzing how people react to the media bias could further our understanding of media influence in sports.

This research has provided valuable insights into the role of media bias in the NBA, highlighting differences in how teams and players are portrayed in the news media. By using sentiment analysis tools and NLP techniques, this study has shown ways in which media narratives can change public perception. The findings highlight the importance of critical media consumption, encouraging fans, players, and organizations to be more aware of potential biases in sports reporting. Overall, this study serves as a foundation for continued exploration into media analysis within professional sports. While this research focuses on the NBA, the implications extend beyond basketball, raising broader questions about fairness, and fair representation in sports media across the world.. As technology advances and media consumption evolves, paying attention to all media narratives will remain important to ensuring balanced and ethical reporting. By bringing attention to the presence of media bias in the NBA, this study contributes to a larger conversation about fairness and bias in sports media. Whether it is through improved reporting practices, or informing the fans better about media, or new tools for detecting bias, the goal is to create a fairer and more accurate portrayal of athletes and teams in the NBA. As the NBA continues to grow, ensuring fair and equitable media representation will be important in maintaining the integrity of the game and the narratives surrounding it.

References

[1] Swati Aggarwal, Tushar Sinha, Yash Kukreti, and Siddarth Shikhar. 2020. Media bias detection and bias short term impact assessment. *Array* 6 (2020), 100025.

[2] Explosion AI. n.d.. spaCy Processing Pipelines. `https://spacy.io/usage/processing-pipelines` Accessed: 2025-02-27.

[3] Alexandra Balahur. 2013. Sentiment analysis in social media texts. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*. 120–128.

[4] Eric Baucom, Azade Sanjari, Xiaozhong Liu, and Miao Chen. 2013. Mirroring the real world in social media: Twitter, geolocation, and sentiment analysis. In *Proceedings of the 2013 international workshop on Mining unstructured big data using natural language processing*. 61–68.

[5] Liang-Chu Chen, Chia-Meng Lee, and Mu-Yen Chen. 2020. Exploration of social media for sentiment analysis using deep learning. *Soft Computing* 24, 11 (2020), 8187–8197.

[6] Katherine J Cramer. 2016. *The politics of resentment: Rural consciousness in Wisconsin and the rise of Scott Walker*. University of Chicago Press.

[7] Zulfadzli Drus and Haliyana Khalid. 2019. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science* 161 (2019), 707–714.

[8] Arne Feddersen, Brad R Humphreys, and Brian P Soebbing. 2018. Sentiment bias in national basketball association betting. *Journal of Sports Economics* 19, 4 (2018), 455–472.

[9] Luciano Floridi and Mariarosaria Taddeo. 2016. What is data ethics? , 20160360 pages.

[10] Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries* 20, 4 (2019), 391–415.

[11] Reiner Jasin. 2024. NBA 2K25 Player Complete Dataset. `https://www.kaggle.com/datasets/reinerjasin/nba-2k25-player-complete-dataset` Accessed on February 26, 2025.

[12] Joseph Kobi. 2024. 1. developing dashboard analytics and visualization tools for effective performance management and continuous process improvement. *International journal of innovative science and research technology* (2024).

[13] Zach Leatherman. n.d.. 11ty: A Simpler Static Site Generator. `https://www.11ty.dev/` Accessed: 2025-02-27.

[14] Nan Li and Desheng Dash Wu. 2010. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision support systems* 48, 2 (2010), 354–368.

[15] Steven Loria. n.d.. TextBlob: Simplified Text Processing. `https://textblob.readthedocs.io/en/dev/` Accessed: 2025-02-27.

[16] NewsAPI. 2025. NewsAPI - Get the latest news from multiple sources. `https://newsapi.org` Accessed: 2025-02-26.

[17] Noor Nashriq Ramly, Fazli Mat Nor, Nurul Haszeli Ahmad, and Mohd Haris Aziz. 2012. Comparative analysis on data visualization for operations dashboard. *International Journal of Information and Education Technology* 2, 4 (2012), 287–290.

[18] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications* 237 (2024), 121641.

[19] Annett Schulze, Fabian Brand, Johanna Geppert, and Gaby-Fleur Böl. 2023. Digital dashboards visualizing public health data: a systematic review. *Frontiers in Public Health* 11 (2023), 999958.

[20] Gayane Sedrakyan, Erik Mannens, and Katrien Verbert. 2019. Guiding the choice of learning dashboard visualizations: Linking dashboard design and data visualization concepts. *Journal of Computer Languages* 50 (2019), 19–38.

[21] Doron Shultziner and Yelena Stukalin. 2021. Distorting the news? The mechanisms of partisan media bias and its effects on news production. *Political Behavior* 43, 1 (2021), 201–222.

[22] Statista. 2024. Total NBA League Revenue 2024. `https://www.statista.com/statistics/193467/total-league-revenue-of-the-nba-since-2005/` Accessed: 2025-01-23.

[23] Fabian Wunderlich and Daniel Memmert. 2020. Innovative approaches in sports science—lexicon-based sentiment analysis as a tool to analyze sports-related Twitter communication. *Applied sciences* 10, 2 (2020), 431.