**NBA Rookie Scouting Report Sentiment Analysis**

Group Name: Analytics Avengers

Group Leader: Tenzin Nargee

Group Members: Kian Putnam, Michael Hijduk

DS4002

2/10/25

**Goal Statement:** Investigate whether the sentiment expressed in 2023 NBA rookie scouting reports is a significant predictor of players’ performance metrics during their rookie season.

**Research Question:** Do the sentiment scores of the 2023 NBA rookie pre-draft scouting reports correlate with the players’ performance metrics in their rookie season?

**Modeling Approach:**

* Scrape text data from various reputable scouting report pages for 2023 NBA Draft [1,2].
* Apply the VADER sentiment analysis package in Python to compute sentiment scores ranging from -1 to 1, where scores between -0.05 and 0.05 are considered neutral.
* Use a combination of correlation tests to assess the relationship between sentiment scores and rookie performance metrics [3,4].
* Define rookie success using measurable criteria such as rookie team awards, minutes played during the season, and win shares.

**Executive Summary:**

This document outlines our data set establishment and analysis plan for examining whether sentiment in 2023 NBA rookie pre-draft scouting reports can predict rookie performance. It details our approach for data scraping, text preprocessing with NLTK, sentiment analysis using VADER, and correlation tests to evaluate relationships between sentiment scores and performance metrics.The goal of this analysis is to investigate the predictive power of sentiment in scouting reports on measurable rookie performance metrics, such as awards, minutes played, and win shares.

**Dataset Establishment Details:**

Goal: The goal of this section is to provide a detailed explanation of the data collected for this analysis, including data sources, preprocessing steps, and exploratory analysis findings. The dataset focuses on 2023 NBA rookies and the sentiment expressed in their scouting reports, which will be correlated with their performance metrics.

Data Set:

Scouting Report data: Web scraped text data of pre-draft scouting reports for all NBA rookies (2023).

Sentiment Scores: The sentiment of each scouting report is quantified using VADER sentiment analysis, producing compound, positive, negative, and neutral sentiment scores.

Performance Metrics: Rookie performance metrics such as minutes played, win shares, player efficiency rating (PER), VORP (Value Over Replacement Player), and rookie awards [5].

Data Dictionary:

| **Column** | **Description** | **Potential Responses** |
| --- | --- | --- |
| Player Name | Name of NBA Player | Victor Wembanyama, Brandon Miller |
| Scouting Report | Full text of pre-draft scouting report | Athletic player with some strengths and some weaknesses |
| Compound Sentiment Score | Overall sentiment score of report | Range from -1 to 1 |
| Positive Sentiment Score | Score that was of positive sentiment | Percentage positive |
| Negative Sentiment Score | Score that was of negative sentiment | Percentage negative |
| Neutral Sentiment Score | Score that was of neutral sentiment | Percentage neutral |
| Player Plus Minus | Player’s overall contribution to team success | Number can be positive or negative depending on scoring while player is on court |
| Player Minutes Played | Total minutes played during rookie season | Total number of minutes player was on court |
| Player PER | Player efficiency rating, for overall efficiency of player | Number can vary, league average is around 15 |
| Player WS | Win shares, showing a player’s contribution to wins | Number showing contribution to a number of games |
| Player VORP | Value over replacement player, showing a player’s overall value | Number quantifying how much better a player is than a replacement |
| Player Awards | Any award received during the rookie season | All rookie first team, rookie of the year |

Questions Explored in EDA:

1. What is the overall distribution of sentiment in scouting reports?

Most compound sentiment scores are clustered near 1.0, suggesting that VADER interprets nearly all reports as overwhelmingly positive. There is very little variation in sentiment across different players.

1. Do scouting reports contain enough variation in sentiment to correlate with performance metrics?

Despite having positive, negative, and neutral sentiment scores, the distribution is heavily skewed toward neutral and positive words. The lack of strong negative language limits the spread of scores, which may affect correlation results.

1. Are there any outliers in sentiment scores?

The variation in compound sentiment scores is minimal. Even players with more critical scouting reports received high compound scores due to the general positivity in scouting report phrasing.

Current Unknowns:

What performance metrics are most correlated with scouting report sentiment? (explored later using correlation tests)

Does sentiment variation (even if small) still correlate with performance metrics?

Are there better sentiment models than VADER for scouting reports? Since VADER is optimized for social media and casual text, it may not be ideal for formal scouting reports [6].

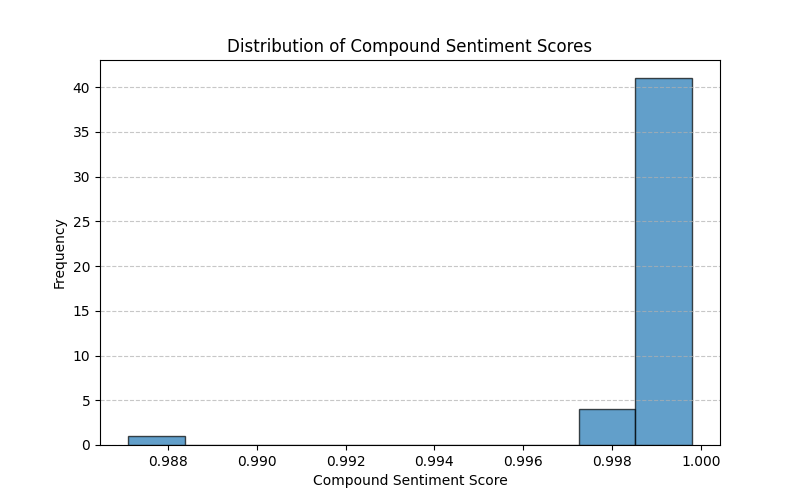
Do sentiment scores vary based on the position or team of the rookie? (could be addressed in the future)

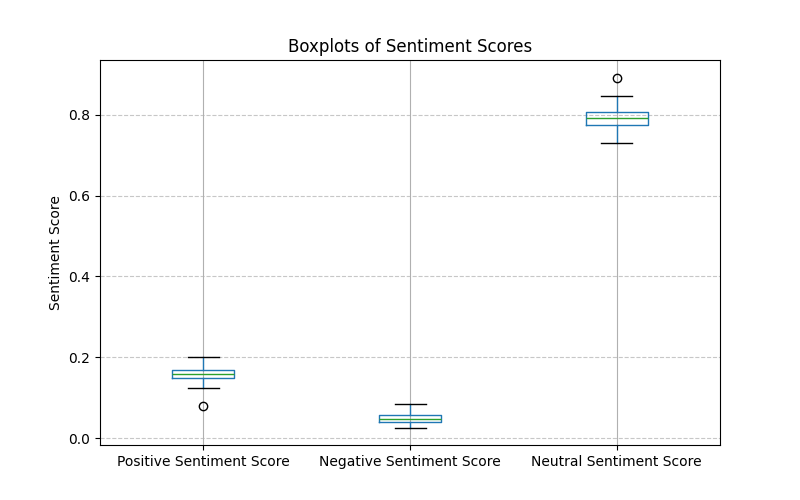
Refinement:

* We are still refining our analysis plan to determine which performance metrics best show the effect of sentiment.
* Exploring alternative sentiment models (e.g., TextBlob, Transformer-based models) to detect subtle variations in tone [7].
* Separating Strengths vs. Weaknesses sections in scouting reports to compare sentiment more effectively.
* Checking correlations despite the low sentiment variation to see if even small sentiment differences impact performance.

Exploratory Plots

* (1) Histogram of Compound Sentiment Scores
* (2) Boxplots of Positive, Negative, and Neutral Sentiment Scores





**Analysis Plan**

#### Goal:

The goal of the analysis is to determine whether the sentiment expressed in the pre-draft scouting reports of NBA rookies can predict their performance metrics during their rookie season.

Graphic for analysis plan:



Methodology and preprocessing:

* Our methodology will start with preprocessing. We will continue cleaning the text to remove unwanted characters and numbers and bring in missing data. We will then utilize NLTK to tokenize the text into individual words. We will also remove the common words in the text that don’t hold meaning.
* Following this we will begin to apply our sentiment analysis using the VADER package. This will give us an overall sentiment score along with a positive, negative, and neutral score. We will focus mostly on the overall sentiment.
* Next is our statistical analysis. We will be performing multiple correlation tests to assess the relationship between the sentiment and the performance of the rookies. This includes a Pearson correlation for minutes and win shares, Spearman correlation for any ranked variables, and Point-Biserial correlation for winning any rookie awards.
* The primary evaluation metric will be derived from the p values. We will be using a p value of less than 0.05 to show a statistically significant relationship with the sentiment and performance.

Specific quantifiable goal:

Achieve correlations between sentiment scores and key rookie performance metrics such as minutes played, win shares, and rookie awards.

**References**

[1] "2023 NBA Draft – Actual Draft Results," NBADraft.net, 2023. [Online]. Available:<https://www.nbadraft.net/actual-draft/?year-mock=2023>. [Accessed: Feb. 5, 2025].

[2] "2023 NBA Draft – ESPN Insider," ESPN, 2023. [Online]. Available:<https://insider.espn.com/nba/draft/rounds/_/season/2023>. [Accessed: Feb. 5, 2025].

[3] "2024 NBA Rookies | Basketball-Reference.com," Basketball-Reference, 2024. [Online]. Available:<https://www.basketball-reference.com/leagues/NBA_2024_rookies.html>. [Accessed: Feb. 5, 2025].

[4] "NBA Player Stats – Traditional," NBA.com, 2024. [Online]. Available:<https://www.nba.com/stats/players/traditional?DraftYear=2023&Season=2023-24&dir=A&sort=GP>. [Accessed: Feb. 5, 2025].

[5] "The Numbers Don't Lie," Samford University, 2023. [Online]. Available:<https://www.samford.edu/sports-analytics/fans/2023/The-Numbers-Dont-Lie>. [Accessed: Feb. 5, 2025].

[6] "nltk.sentiment.vader Module Documentation," NLTK, 2024. [Online]. Available:<https://www.nltk.org/api/nltk.sentiment.vader.html>. [Accessed: Feb. 7, 2025].

[7] A. S. M. Sultan, M. A. Rahman, and M. S. Hossain, "TextBlob and BiLSTM for Sentiment Analysis Toward COVID-19 Tweets," *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, Sylhet, Bangladesh, 2022, pp. 1-6. [Online]. Available:<https://ieeexplore.ieee.org/document/9736380/>. [Accessed: Feb. 10, 2025].