CIS 400 Term Project

# Sentiment Analysis Regarding the Comparison between Performance of Basketball Players/Teams and Twitter Comments


**Team Members:**

| | |
|---|---|
| Zongxiu Wu | (SUID: 52275-8233) |
| Rick Li | (SUID: 25849-8271) |
| Runzhi Ma | (SUID: 93641-8751) |
| Yiheng Lu | (SUID: 44623-9041) |
| Matthew Keenan | (SUID: 28168-2209) |




Social Media and Data Mining (CIS - 400) Syracuse University
New York
May 2022

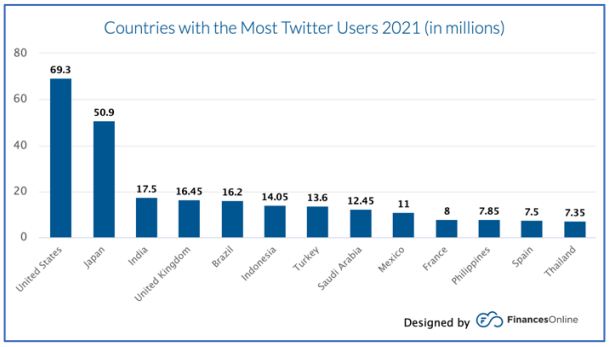
# Table of Contents

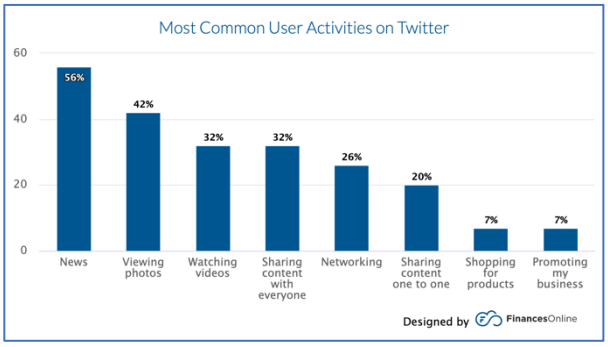
# I. Abstract

Twitter is a social media service provider that hosts a platform for users to send microblogs named tweets. Except for posting new tweets, user interaction on Twitter are likes, comments, shares, and retweets. There are highly influential accounts such as celebrities posting tweets every day and receiving a lot of retweets and comments, including the famous basketball players and official accounts for basketball teams. While the highly influential accounts can make announcements, the public can also post comments on how they view the influential individuals. In this paper, we will perform sentiment analysis on people's tweets that are related to certain basketball players and teams. We will also compare the analysis results with the players' and teams' actual performance in current and previous seasons.

# II. Introduction

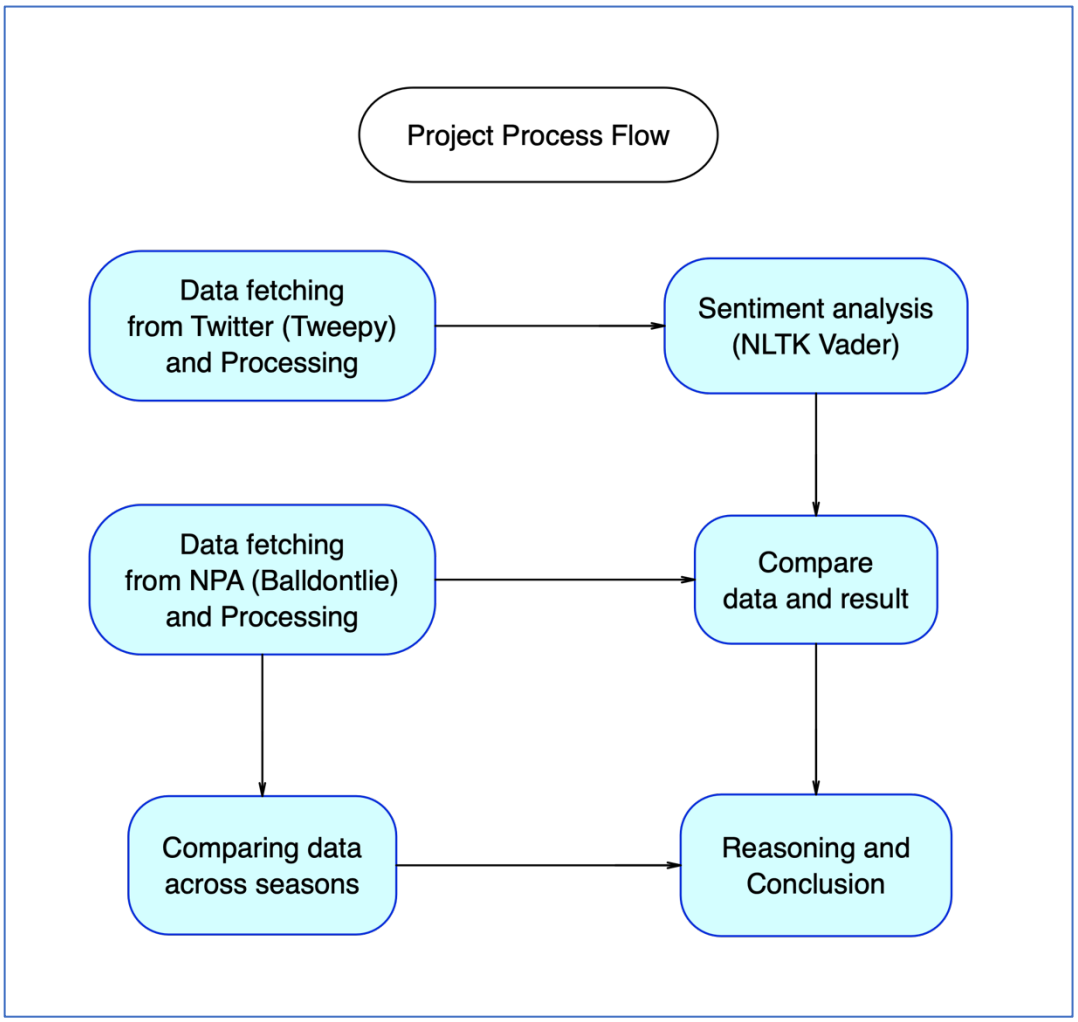
As a US company, the USA is the country that is particularly popular for Twitter users. By 2021, there were more than 69 million users in the United States and 35 million daily active users. Even though Twitter is not as popular as Facebook, there are still approximately 500 million tweets sent a day, and more than 32% are related to people sharing content publicly (FinanceOnline, 2022).




The content shared by Twitter users contains various categories, one of them is people's comments on basketball players and teams. The tweet in the format of "@" a specific basketball player/team's account followed by comments can be commonly found. Those tweets contain Twitter users' opinions, expressions of emotion, and criticism. However, because people are not restricted with their speech, their tweets can also contain false information or non-sense words. Therefore, to find out whether people are expressing their feeling on Twitter in a reasonable way, it is worth trying to fetch out those comments, investigate their emotions, and compare them with the players/teams' actual performance in the current season.

## III. Process Flow



# IV. Balldontlie API Introduction

For the goal of conducting an analysis based on Player/Team's performances each year. The Ball-don't-lie Application Programming Interface was used to help fetch data regarding NBA status.

- No email required
- No API key required
- The API contains all NBA Player/Team's data from seasons 1979-current
- Live(ish) game stats are available (updated every 10 minutes)
- Rate limit of 60 requests per minute

Link to the API website -- https://www.balldontlie.io/

### Fetching NBA Player/Team data using Balldontlie API

Requirements for using Balldontlie API:

- It is a free API source project posted by the Twitter account "Knockout Ned" and its team members with neither email nor API key required
- A computer with an internet connection

### Fetching NBA Player's data

One shortcoming of using Balldontlie API is that it only stores player info from 1978 to current, meaning that once looking for a player who started playing before 1978 could result in missing data.

However, Balldontlie API is overall powerful, it provides functionalities including ***get_all_player_data***, ***get_a_specific_player_data***, ***get_all_player_status***, and ***get_player_season_average***. With a variety of parameter imputations, the functions could draw conclusively data sets with exhaustive details.

### Fetching NBA Team's data

In addition to fetching NBA player data, the Balldontlie API supports functionality for fetching NBA team data. This part of the API can return team data through the "***get_all_teams***" and "***get_a_specific_team***" options available. The only requirement to use the "***get_a_specific_team***" options is the unique id associated with the team. Given that there are 32 NBA teams, a team's unique id will be a number between 1 and 32.

**API Methods Used**

The Balldontlie API supports multiple embedded functions with implementations of different parameters.

1. *Get a Specific Player*
    - For fetching the player's information, the API requests a specific player's name to start, returning the player's information, including the player's id number, age, height, weight, and so on data.
    - Since the data itself is stored in dictionary form, the player's id can be easily extracted and saved as a variable, which will be later passed down to the second API call to retrieve yearly performances.

2. *Get Player's Season Averages*
    - To further abstract yearly performances, the Player's id is necessary with another yearly average request to the API itself, returning all essential data including the player's average points achieved, number of assistants, rebounds, and turnovers from the current season to all the way back up to season 2000.
    - Finally, the data will be recorded in dictionary form both displaying to the command line and gets printed into a text file name varied by the player's name.

3. *Get All Teams*
    - In order to verify that the input of a requested team is that of a valid NBA team, this API method was called to aggregate information on all 32 NBA teams
    - For each team returned, data such as name, full name, city, and abbreviation were provided, and this information was used against the given input to see if the input matched any of the NBA teams

4. *Get a Specific Team*
    - At some points, the NBA team whose data was needed was known, meaning that returning data on just the one team would be more efficient than returning data on all 32 NBA teams
    - The one additional piece of data required was the team's unique id, and this was retrieved and stored in the previous step in which the requested team was verified to be a valid NBA team
    - With the saved team id appended to the request, the "Get a Specific Team" method returned relevant data on just the one team requested

5. *Get All Games*
    - Because the Balldontlie API does not provide a method for returning NBA team season averages, the “Get All Games” method that is provided was used to return this data by collecting data from each game in a season for the required team
    - Data such as total games, total wins, total points, points per game, as well as season win percentage were calculated after aggregating data from each game in the season, and this data was saved into a dictionary format, displayed to the command line, and printed to a text file

**Where does the data go to?**

To prevent situations when data is needed for implementation, comparison, and return, abstracting the data requires running the whole program first, which engenders a lot of issues including unnecessary running time waste and rate limitations. Thus, after implementing Balldontlie functions, their corresponding data will be stored in dictionary form and get printed to both the command line and a text file name varied by the player or team name. For example, in case fetching a player’s info, his average data per season will be generated into a file named "Player-- * data.txt", in which the symbol * represents the name of the player. Similarly, while storing a team’s data, the file will be named "Team-- * data.txt".

Thus, with three players and three teams getting analyzed, this project ends up creating six data files in total, which are:

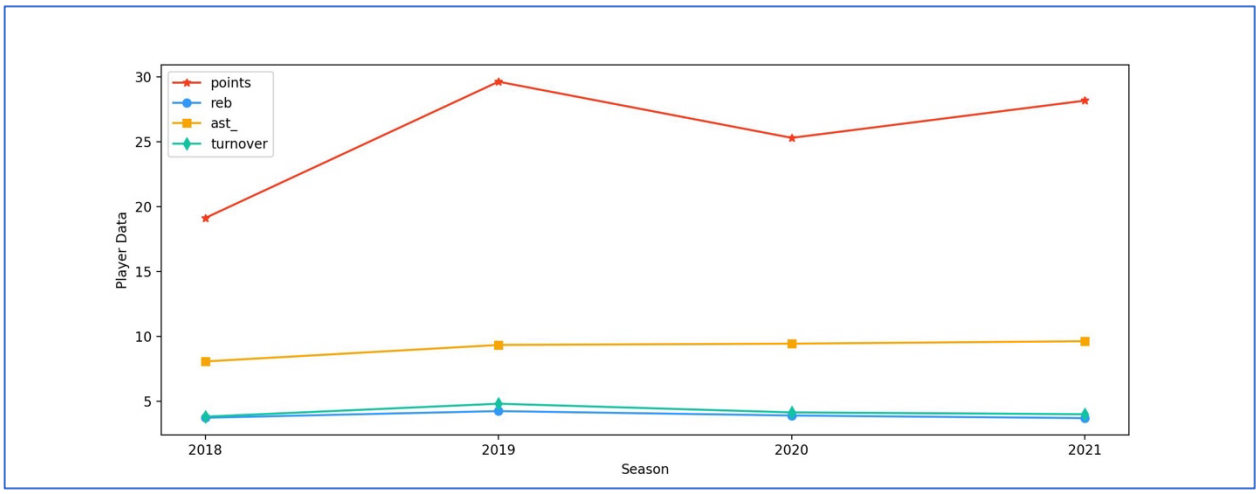
1. "Player-- Trae_Young data.txt"
2. "Player-- Lebron_James data.txt"
3. "Player-- James_Harden data.txt"
4. "Team-- Atlanta Hawks data.txt"
5. "Team-- Brooklyn Nets data.txt"
6. "Team-- Los Angeles Lakers data.txt"

Besides, the data will be transposed to the next procedure of generating graphics for comparison analysis.

## Player Table: Average Performances per Season

**Player:** ***Trae Young***

| Season | PTS | AST | REB | STL | BLK | FTA | FTM | FGA | FGM | TO | GP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ***2018*** | 19.12 | 8.06 | 3.72 | 0.86 | 0.19 | 5.11 | 4.23 | 15.51 | 6.48 | 3.8 | 81 |
| ***2019*** | 29.63 | 9.33 | 4.23 | 1.08 | 0.13 | 9.32 | 8.02 | 20.82 | 9.1 | 4.8 | 60 |
| ***2020*** | 25.3 | 9.43 | 3.89 | 0.84 | 0.17 | 8.67 | 7.68 | 17.65 | 7.73 | 4.13 | 63 |
| ***2021*** | 28.17 | 9.62 | 3.69 | 0.94 | 0.09 | 7.22 | 6.53 | 20.25 | 9.3 | 3.99 | 77 |




Range from 2018 to 2021

Points:

Bimodal distribution, with one local peak at year 2019 and one low peak at year 2020

Rebounds, Assists, Turnovers:

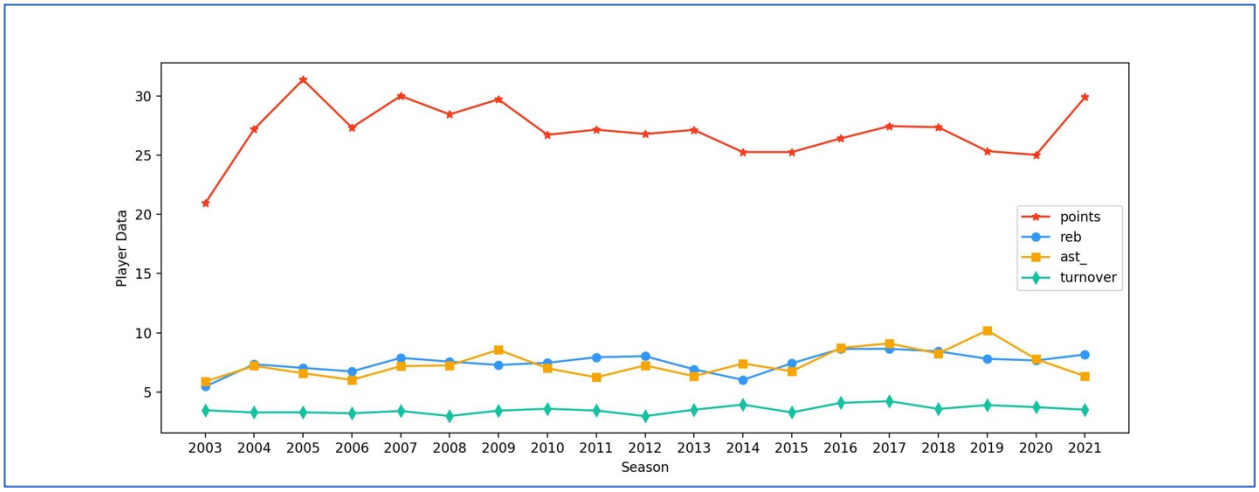
An overall uniform distribution, with no local peak over four years term.

With not many statistics driven for Trae's performances, his scores are averagely distributed

**Player:** ***Lebron James***

| Season | PTS | AST | REB | STL | BLK | FTA | FTM | FGA | FGM | TO | GP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **2003** | 20.94 | 5.89 | 5.47 | 1.65 | 0.73 | 5.82 | 4.39 | 18.89 | 7.87 | 3.46 | 79 |
| **2004** | 27.19 | 7.21 | 7.35 | 2.21 | 0.65 | 7.95 | 5.96 | 21.05 | 9.94 | 3.28 | 80 |
| **2005** | 31.37 | 6.59 | 7.04 | 1.56 | 0.84 | 10.3 | 7.61 | 23.08 | 11.08 | 3.29 | 79 |
| **2006** | 27.33 | 6.03 | 6.74 | 1.6 | 0.71 | 8.99 | 6.27 | 20.78 | 9.9 | 3.21 | 78 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **2007** | 30.0 | 7.19 | 7.89 | 1.84 | 1.08 | 10.28 | 7.32 | 21.89 | 10.59 | 3.4 | 75 |
| **2008** | 28.44 | 7.25 | 7.57 | 1.69 | 1.15 | 9.41 | 7.33 | 19.91 | 9.74 | 2.98 | 81 |
| **2009** | 29.71 | 8.57 | 7.29 | 1.64 | 1.01 | 10.17 | 7.8 | 20.11 | 10.11 | 3.43 | 76 |
| **2010** | 26.72 | 7.01 | 7.47 | 1.57 | 0.63 | 8.39 | 6.37 | 18.8 | 9.59 | 3.59 | 79 |
| **2011** | 27.15 | 6.24 | 7.94 | 1.85 | 0.81 | 8.1 | 6.24 | 18.85 | 10.02 | 3.44 | 62 |
| **2012** | 26.79 | 7.25 | 8.03 | 1.7 | 0.88 | 7.04 | 5.3 | 17.82 | 10.07 | 2.97 | 76 |
| **2013** | 27.13 | 6.34 | 6.92 | 1.57 | 0.34 | 7.6 | 5.7 | 17.57 | 9.96 | 3.51 | 77 |
| **2014** | 25.26 | 7.41 | 6.03 | 1.58 | 0.71 | 7.65 | 5.43 | 18.54 | 9.04 | 3.94 | 69 |
| **2015** | 25.26 | 6.76 | 7.43 | 1.37 | 0.64 | 6.46 | 4.72 | 18.63 | 9.7 | 3.28 | 76 |
| **2016** | 26.41 | 8.73 | 8.64 | 1.24 | 0.59 | 7.18 | 4.84 | 18.16 | 9.95 | 4.09 | 74 |
| **2017** | 27.45 | 9.11 | 8.65 | 1.41 | 0.87 | 6.48 | 4.73 | 19.27 | 10.45 | 4.23 | 82 |
| **2018** | 27.36 | 8.25 | 8.44 | 1.31 | 0.6 | 7.6 | 5.05 | 19.91 | 10.15 | 3.58 | 55 |
| **2019** | 25.34 | 10.21 | 7.81 | 1.16 | 0.55 | 5.69 | 3.94 | 19.45 | 9.6 | 3.9 | 67 |
| **2020** | 25.02 | 7.78 | 7.67 | 1.07 | 0.56 | 5.67 | 3.96 | 18.29 | 9.38 | 3.73 | 45 |
| **2021** | 29.89 | 6.34 | 8.17 | 1.36 | 1.04 | 5.81 | 4.42 | 21.72 | 11.34 | 3.51 | 53 |




Range from 2003 to 2021

Points:

Right skewed distribution, with a local high peak in the 2005 year and low peak in the 2021 year.

One turning shift in the year 2021.

Assists:

Multimodal distribution, one high peak in the year 2019.

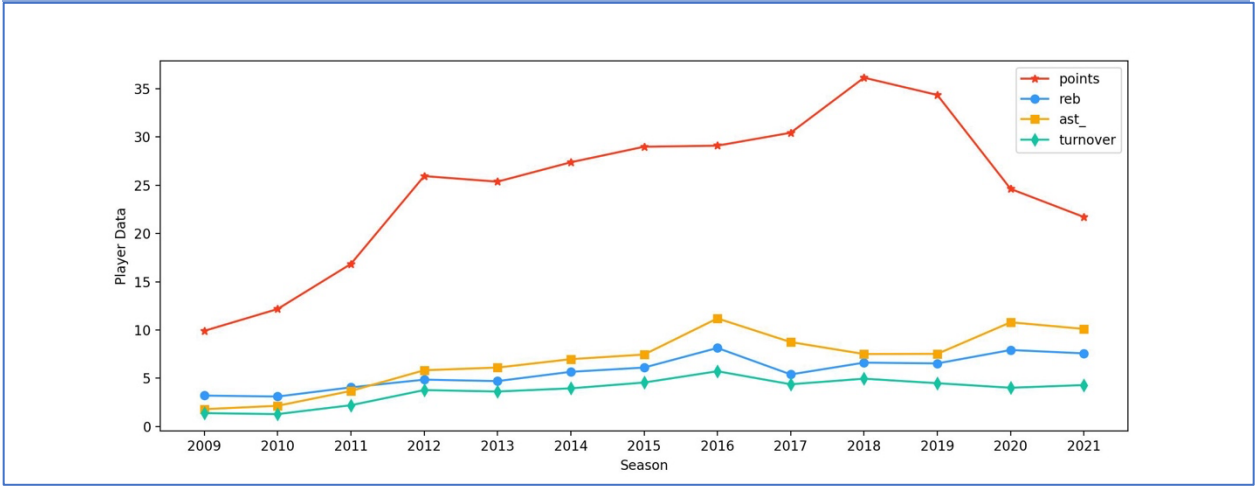
Rebounds, Turnovers:

An overall uniform distribution, with no local peak over four years terms.

All Lebron's statistics are placed at an averagely high level, demonstrating his amazing performances in the NBA league.

**Player:** ***James Harden***

| Season | PTS | AST | REB | STL | BLK | FTA | FTM | FGA | FGM | TO | GP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ***2009*** | 9.91 | 1.8 | 3.21 | 1.05 | 0.26 | 3.16 | 2.55 | 7.61 | 3.07 | 1.39 | 76 |
| ***2010*** | 12.17 | 2.15 | 3.11 | 1.12 | 0.29 | 4.18 | 3.52 | 8.34 | 3.63 | 1.29 | 82 |
| ***2011*** | 16.84 | 3.69 | 4.06 | 1.0 | 0.24 | 5.95 | 5.03 | 10.15 | 4.98 | 2.21 | 62 |
| ***2012*** | 25.94 | 5.83 | 4.86 | 1.82 | 0.49 | 10.15 | 8.64 | 17.14 | 7.5 | 3.78 | 78 |
| ***2013*** | 25.36 | 6.11 | 4.71 | 1.58 | 0.4 | 9.11 | 7.89 | 16.51 | 7.52 | 3.63 | 73 |
| ***2014*** | 27.37 | 6.98 | 5.67 | 1.9 | 0.74 | 10.17 | 8.83 | 18.15 | 7.99 | 3.96 | 81 |
| ***2015*** | 28.98 | 7.46 | 6.11 | 1.7 | 0.62 | 10.21 | 8.78 | 19.72 | 8.66 | 4.56 | 82 |
| ***2016*** | 29.09 | 11.2 | 8.14 | 1.49 | 0.47 | 10.88 | 9.21 | 18.93 | 8.32 | 5.73 | 81 |
| ***2017*** | 30.43 | 8.75 | 5.4 | 1.75 | 0.69 | 10.1 | 8.67 | 20.13 | 9.04 | 4.38 | 72 |
| ***2018*** | 36.13 | 7.51 | 6.62 | 2.04 | 0.73 | 11.0 | 9.67 | 24.47 | 10.81 | 4.96 | 78 |
| ***2019*** | 34.34 | 7.53 | 6.54 | 1.84 | 0.88 | 11.76 | 10.18 | 22.26 | 9.88 | 4.49 | 68 |
| ***2020*** | 24.61 | 10.8 | 7.93 | 1.2 | 0.75 | 7.34 | 6.32 | 16.68 | 7.77 | 4.02 | 44 |
| ***2021*** | 21.7 | 10.11 | 7.58 | 1.24 | 0.55 | 8.12 | 7.12 | 15.03 | 6.17 | 4.3 | 66 |




Range from 2009 to 2021

Points:

Left skewed distribution, with two local high peaks at 2012 year and 2018 year.

Assists, Rebounds, Turnovers:

Left skewed distribution, with local high peak at year 2016.

James Harden has been improving since he started playing in NBA, his overall turnovers generate less association with his points graph. James met his biggest change in 2018, according to the graph we can conclude that James chose to be less aggressive about winning points but paid more attention to making assistants for his teammates.

## Player Efficiency Rating Analysis

### How to conduct an analysis for players?

Calculating a player's performance could be arduous, the official NBA league has developed a set of formulas to conduct evaluation and comparisons between players. However, those equations are complicated and involve countless variables.

One example is "The Player Efficiency Rating" (PER). PER is a per-minute rating developed by ESPN.com columnist John Hollinger. In John's words, "The PER sums up all a player's positive accomplishments, subtracts the negative accomplishments, and returns a per-minute rating of a player's performance."

Yet, Thanks to Martin Manley, a Kansas City sports reporter and statistician, who conceived a simpler version of calculating a player's PER, a player's accomplishment in a year can be effortlessly calculated.

The Simple PER formula was stated as:

**SPER = (PTS + REB + AST + STL + BLK – (FTA - FTM) – (FGA - FGM) - TO) / GP**

In which,

**PTS** = Points, **REB** = Rebounds, **AST** = Assists, **STL** = Steals, **BLK** = Blocks

**FTA**, **FTM** = Free Throws Attempted, Free Throws Made

**FGA**, **FGM** = Field Goals Attempted, Field Goals Made

**TO** = Turnovers, **GP** = Games Played

From the above formula we compute the season average SPER for the three players:

Trae Young, Lebron James, and James Harden

**SPER from Season 2020 to 2021**

| Player | SPER in Season 2020 | SPER in Season 2020 | SPER in Season 2021 | Difference | Difference |
|---|---|---|---|---|---|
| **Trae Young** | 24.5900 | 24.5900 | 26.8800 | +9.31% | +9.31% |
| **Lebron James** | 27.7500 | 27.7500 | 31.5200 | +13.59% | +13.59% |
| **James Harden** | 31.3400 | 31.3400 | 27.0200 | -13.78% | -13.78% |

In which Lebron James performed the greatest advance on his SPER value from season 2020 to 2021 with a percentage of 13.59%, secondly Trae acquired and 9.31% advanced on his SPER. On the other hand, James' SPER got lowered by 13.78%.

Despite any advance or retreat from the player, from statistics we conclude that in season 2021 Lebron James plays the best out of the three players, followed by James Harden then finally Trae Young

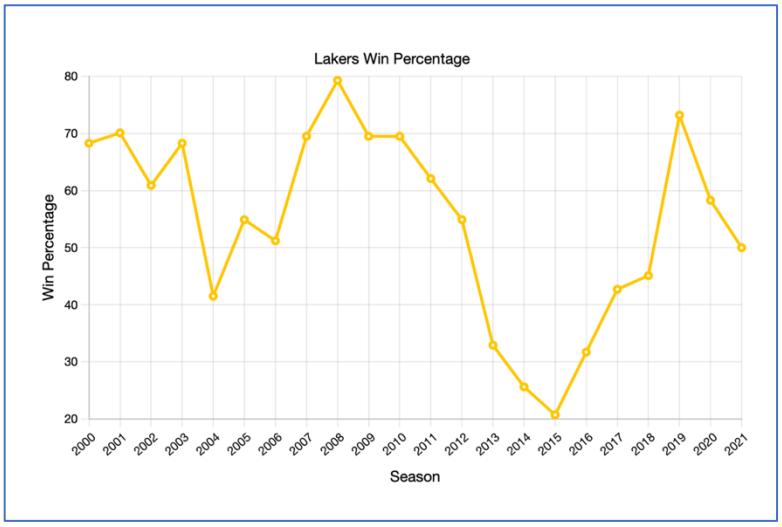
# Team Table: Average Performances per Season

**Team:** ***Los Angeles Lakers***

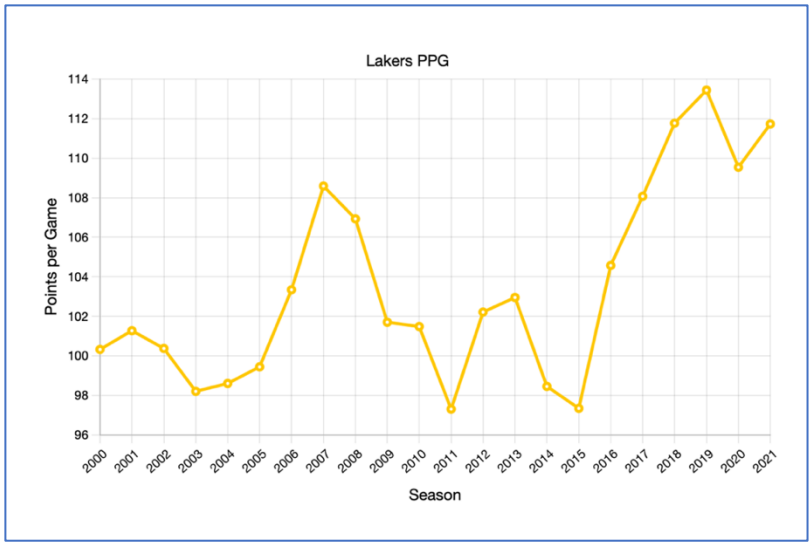
| Season | Total Wins | Tot. Games | Win Pct. | Total Points | PPG |
|---|---|---|---|---|---|
| **2000** | 56 | 82 | 68.3 | 8226 | 100.32 |
| **2001** | 58 | 82 | 70.1 | 8304 | 101.27 |
| **2002** | 50 | 82 | 60.9 | 8230 | 100.37 |
| **2003** | 56 | 82 | 68.3 | 8052 | 98.2 |
| **2004** | 34 | 82 | 41.5 | 8085 | 98.6 |
| **2005** | 45 | 82 | 54.9 | 8154 | 99.44 |
| **2006** | 42 | 82 | 51.2 | 8474 | 103.34 |
| **2007** | 57 | 82 | 69.5 | 8904 | 108.59 |
| **2008** | 65 | 82 | 79.3 | 8768 | 106.93 |
| **2009** | 57 | 82 | 69.5 | 8339 | 101.7 |
| **2010** | 57 | 82 | 69.5 | 8321 | 101.48 |
| **2011** | 41 | 82 | 62.1 | 6422 | 97.3 |
| **2012** | 45 | 82 | 54.9 | 8381 | 102.21 |
| **2013** | 27 | 82 | 32.9 | 8442 | 102.95 |
| **2014** | 21 | 82 | 25.6 | 8073 | 98.45 |
| **2015** | 17 | 82 | 20.7 | 7982 | 97.34 |
| **2016** | 26 | 82 | 31.7 | 8575 | 104.57 |

| | | | | | |
|---|---|---|---|---|---|
| **2017** | 35 | 82 | 42.7 | 8862 | 108.07 |
| **2018** | 37 | 82 | 45.1 | 9165 | 111.77 |
| **2019** | 52 | 71 | 73.2 | 8054 | 113.44 |
| **2020** | 42 | 72 | 58.3 | 7887 | 109.54 |
| **2021** | 44 | 88 | 50.0 | 9832 | 111.73 |

Lakers Performance Progression 2000 to 2021




In terms of win percentages since 2000, the Lakers have generally remained in the 40%-80% range, with the exception of 2013-2016. Since 2016, they have been back in the win percentage that they have consistently reached over this 21-year span, leaving the sentiment regarding the Lakers to likely to be higher than it had been between 2013 and 2016.



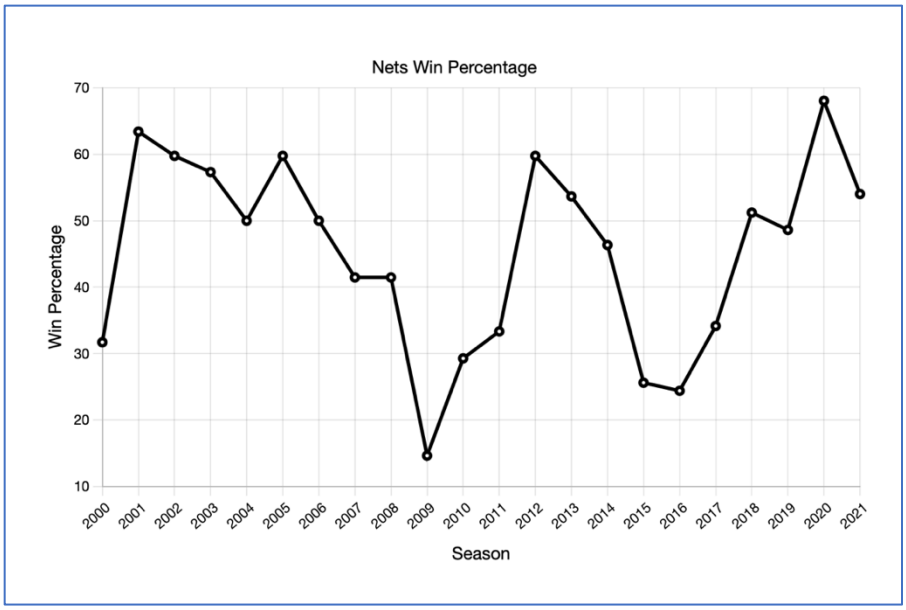

When looking at both the points per game and win percentage in 2015, it shows that both represent the lowest ever for the Lakers since 2000. The data shows that comparing an NBA team's average points per game for a season can have some correlation with their overall win percentage. Excluding the years 2012 and 2013 from the "Points Per Game" graph, similar patterns can be seen in both graphs. There is a consistent increase in both graphs between 2004-2008, and a general decrease from 2008-2011. In addition, the data has a steep increase in both points per game and win percentage starting in 2015.

**Team: *Brooklyn Nets***

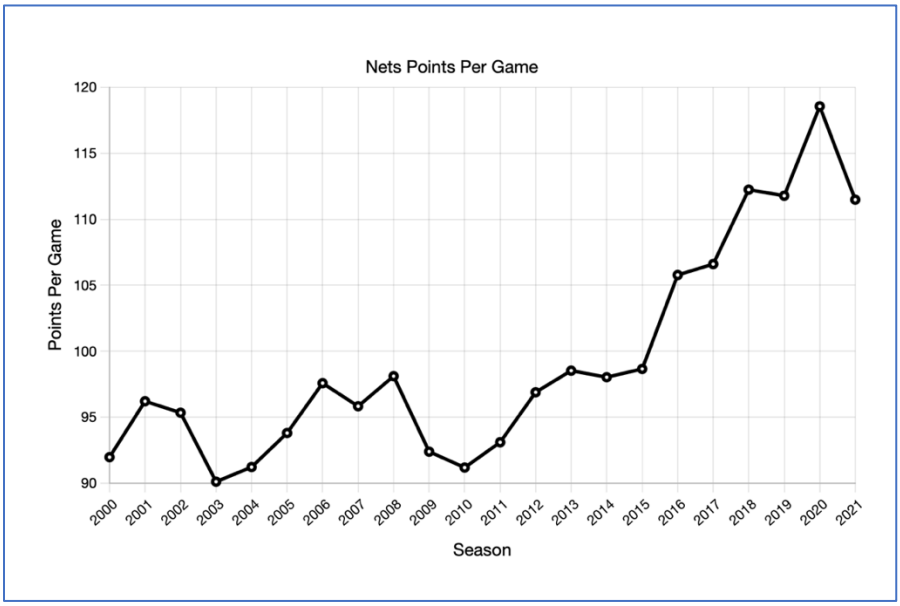
| Season | Total Wins | Tot. Games | Win Pct. | Total Points | PPG |
|---|---|---|---|---|---|
| **2000** | 26 | 82 | 31.7 | 7541 | 91.96 |

| | | | | | |
|---|---|---|---|---|---|
| **2001** | 52 | 82 | 63.41 | 7888 | 96.2 |
| **2002** | 49 | 82 | 59.76 | 7817 | 95.33 |
| **2003** | 47 | 82 | 57.32 | 7388 | 90.1 |
| **2004** | 41 | 82 | 50 | 7480 | 91.21 |
| **2005** | 49 | 82 | 59.76 | 7691 | 93.79 |
| **2006** | 41 | 82 | 50 | 8001 | 97.57 |
| **2007** | 34 | 82 | 41.46 | 7857 | 95.82 |
| **2008** | 34 | 82 | 41.46 | 8044 | 98.1 |
| **2009** | 12 | 82 | 14.63 | 7575 | 92.38 |
| **2010** | 24 | 82 | 29.27 | 7722 | 91.17 |
| **2011** | 22 | 82 | 33.33 | 6144 | 93.09 |
| **2012** | 49 | 82 | 59.76 | 7944 | 96.88 |
| **2013** | 44 | 82 | 53.66 | 8079 | 98.52 |
| **2014** | 38 | 82 | 46.34 | 8038 | 98.02 |
| **2015** | 21 | 82 | 25.61 | 8089 | 98.65 |
| **2016** | 20 | 82 | 24.39 | 8673 | 105.77 |
| **2017** | 28 | 82 | 34.15 | 8741 | 106.6 |
| **2018** | 42 | 82 | 51.22 | 9204 | 112.24 |
| **2019** | 35 | 71 | 48.61 | 8048 | 111.78 |
| **2020** | 49 | 72 | 68.05 | 8537 | 118.56 |
| **2021** | 47 | 87 | 54.02 | 9699 | 111.48 |

Nets Performance Progression 2000-2021




Compared to the Lakers' data, the Nets' performance is worse both in terms of games won and much less consistent year to year. Looking at the graph overall, the Nets generally stay within the 30%-70% range. Their inconsistency comes in the sharp increases and decreases in wins, especially from 2008 to 2012. These sharp differences in performance can leave people with a certain level of uncertainty prior to a season.

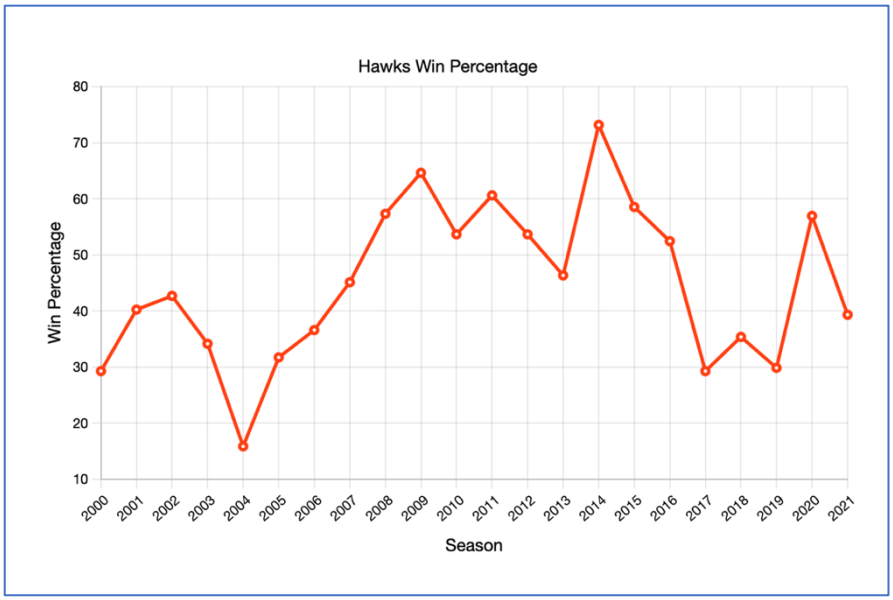



The Nets' points per game data are a good example of how correlation does not equal causation. The Lakers' points per game seemed to have a high correlation with their win percentage, but this is not the case with the Nets. Their points per game from 2000 to 2015 stay generally a similar year to year in the range of 90 to 100 points per game, followed by a sharp spike from 2015 to 2020. This does not exactly follow their win percentages, which appeared much more erratic and inconsistent year to year.

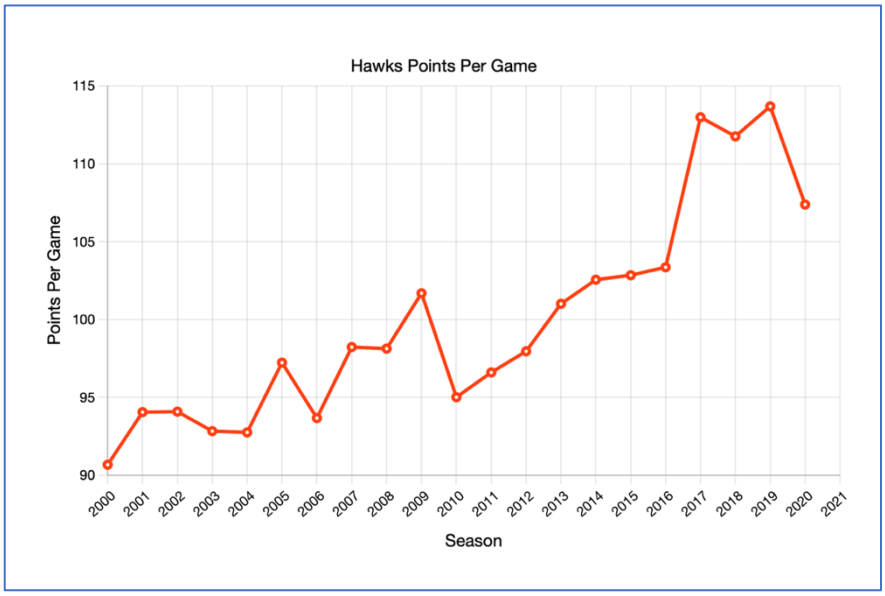
Table for the Atlanta Hawks

| Season | Total Wins | Tot. Games | Win Pct. | Total Points | PPG |
|---|---|---|---|---|---|
| **2000** | 24 | 82 | 29.27 | 7434 | 90.66 |
| **2001** | 33 | 82 | 40.24 | 7711 | 94.04 |
| **2002** | 35 | 82 | 42.68 | 7714 | 94.07 |
| **2003** | 28 | 82 | 34.15 | 7611 | 92.82 |
| **2004** | 13 | 82 | 15.85 | 7605 | 92.74 |
| **2005** | 26 | 82 | 31.71 | 7972 | 97.22 |
| **2006** | 30 | 82 | 36.59 | 7680 | 93.66 |
| **2007** | 37 | 82 | 45.12 | 8054 | 98.22 |
| **2008** | 47 | 82 | 57.32 | 8046 | 98.12 |
| **2009** | 53 | 82 | 64.63 | 8338 | 101.68 |
| **2010** | 44 | 82 | 53.66 | 7790 | 95 |
| **2011** | 40 | 82 | 60.61 | 6375 | 96.59 |
| **2012** | 44 | 82 | 53.66 | 8032 | 97.95 |
| **2013** | 38 | 82 | 46.34 | 8282 | 101 |
| **2014** | 60 | 82 | 73.17 | 8409 | 102.55 |
| **2015** | 48 | 82 | 58.54 | 8433 | 102.84 |
| **2016** | 43 | 82 | 52.44 | 8459 | 103.16 |
| **2017** | 24 | 82 | 29.27 | 8475 | 103.35 |
| **2018** | 29 | 82 | 35.37 | 9265 | 112.99 |
| **2019** | 20 | 71 | 29.85 | 7488 | 111.76 |
| **2020** | 41 | 72 | 56.94 | 8186 | 113.69 |
| **2021** | 35 | 89 | 39.32 | 9557 | 107.38 |

Hawks Performance Progression 2000-2021




While the Hawks' win percentages from 2000-2021 do vary over time, the overall progression of performance is fairly gradual. Overall, they appear to go on winning and losing streaks. From 2004 to 2009, they gradually go up in win percentage, and from both 2009 to 2013 and 2014 to 2017 there is a decrease in win percentage. However, the years between 2008 and 2020 have seen a significant number of seasons at or above the 50% win, likely leaving people with an overall positive and hopeful outlook for the Hawks.




Similar to the Nets, the Hawks' points per game are not indicative of their win percentage over the years. In fact, from 2000 to 2021, their points per game gradually increase nearly the entire time. With the exception of a few years, each season either matches or is larger than the previous season.

# V. Fetching tweets data

## Introduction

In order to analyze if fans' attitudes to a player or a team are positive or negative. We need to come up with a program. It should be able to fetch more than the most recent 5000 tweets which contain a specific NBA player's screen name or an NBA team's screen name. And

then, all the data about should be saved in a CSV file. The players we decided to analyze are Lebron James(@KingJames), James Harden(@JHarden13), and Trae Young(@TheTraeYoung). The teams we decided to analyze are the Los Angeles Lakers(@Lakers), the Brooklyn Nets(@BrooklynNets), and the Atlanta Hawks(@ATLHawks).

**Tweepy API**

The Tweepy API class provides access to the entire twitter RESTful API methods. We can get different kinds of information relating to tweets, followers or other things by giving it different parameters and it would return the information that we need.

**API Methods Used**

The main method I used to fetch the tweets is the "API.search_tweets(kw,max_id=min_id,result_type='recent',count=100)" This method can be used to fetch tweets by giving it different parameters. Kw is the keywords such as "@KingJames". Max_id means it would only return the tweets which have an ID less than or equal to this ID. Result type means that it would return the most recent tweets. The counts mean the number of tweets it would try to return on each page.

**Time Limitation**

One of the difficulties we encounter is the time limit since the API. The search method can only fetch the tweets for the last 7 days. The method we used to solve this problem is running the program at different times. For example, if I run the program once in 2022.4.17, it would be able to fetch all tweets which contain the player's screen name from 4.11 to 4.17. Meanwhile, if I run it again in 4.22, I can get all data that contains the players' screennames from 4.17 to 4.22. Then I just need to combine them and add all tweets into one CSV file.

**Cancel repetition**

When adding tweets to an existing CSV file, one problem that needs to be considered is eliminating the repetition. For example, if we run for the first time in 4.17 and run for the second time in 4.18, tweets that are fetched from 4.12 to 4.17 would appear two times. So, when we want to add new tweets to the existing file, we need to traverse the existing CSV file first. If it already exists, we do not need to add it to the CSV file. If it is not existing in the CSV file already, we add it to the CSV file.

# VI. Sentiment Analysis

### What is sentiment analysis?

Sentiment analysis, by definition, is "the field of study that analyzes people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text. The entities can be products, services, organizations, individuals, events, issues, or topics" (Liu, 2015, p.1). In this paper, NLTK (natural language toolkit) is the API used for performing sentiment analysis. The analyzer provided by NLTK can iterate through every word of people's posted text, assign a score for each word in different layers, and calculate the overall score. The score consists of four objects: compound, positive, negative, and neutral. The compound ranges from -1 to 1, closer to -1 refers to a more negative attitude the tweet contains. In reverse, the closer to 1, the more positive altitude the tweet contains.

### Why use sentiment analysis?

Sentiment analysis will generate a compound indicator for each tweet fetched. For tweets that are related to a specific player/team, sum up the compound number for each tweet and average the sum. The value indicates an overall positive or negative view based on the tweets. Knowing the attitude value, we can compare it with players/teams' performance in the current season and find the correlation.

### Methodology Used

#### *csv reader*

CSV refers to the common separated values. The *csv* module enables users to read and write data that is stored in CSV format. CSV is a preferred format compared to Excel for better compatibility with Python. The *reader* reads into the csv file that stores all tweets from different users on the sample players.

#### *nltk.sentiment.vader*

Vader (Valence Aware Dictionary for Sentiment Reasoning) "is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion" (Beri, 2020). Vader accepts lexical expressions and decides the emotion of such expressions by assigning sentiment scores. For example, "like" as a word will be signed with a positive sentiment score, but "did not like/dislike" will refer to a negative sentiment score.

During the process of evaluating tweets, Vader effectively discovered the sentiment of each comment on teams/players. It can correctly reflect the positive or negative extent of a tweet.

**Results and Finding**

We executed our project on three players and three teams. The sample players are Trae Young, Lebron James, and James Harden. The sample teams are Atlanta Hawks, Brooklyn Nets, and Los Angeles Lakers. The following 6 tables display the most positive and most negative tweets of each player/team.

**Player: *Trae Young***

| User ID | Tweet | Compound | Positive | Negative |
|---|---|---|---|---|
| **1517520779461767169** | '@TheTraeYoung PLEASE WIN' | 0.8208 | 0.883 | 0.0 |
| **1515887386848378880** | '@TheTraeYoung No pressure' | -0.5267 | 0.0 | 0.815 |

**Player: *Lebron James***

| User ID | Tweet | Compound | Positive | Negative |
|---|---|---|---|---|
| **1510870214686044160** | '@KingJames THANK GOD.' | 0.7241 | 0.858 | 0.0 |
| **1515836148895125516** | '@LakersLead @KingJames HELL NO!' | -0.8611 | 0.0 | 0.811 |

**Player: *James Harden***

| User ID | Tweet | Compound | Positive | Negative |
|---|---|---|---|---|
| **1516100858408173570** | '@JHarden13 Good morning, Gorgeous!' | 0.8016 | 0.782 | 0.0 |
| **1511775520857464835** | '@JHarden13 SICK' | -0.6166 | 0.0 | 0.801 |

**Team: *Atlanta Hawks***

| User ID | Tweet | Compound | Positive | Negative |
|---|---|---|---|---|
| **1518956906215264258** | '@ATLHawks better win tonight' | 0.7717 | 0.77 | 0.0 |
| **1518579884003676160** | '@ATLHawks No shit' | -0.7003 | 0.0 | 0.853 |

**Team: *Brooklyn Nets***

| User ID | Tweet | Compound | Positive | Negative |
|---|---|---|---|---|
| **1519493370858684417** | '@BrooklynNets lmao' | 0.5994 | 0.796 | 0.0 |
| **1519350854515675137** | '@BrooklynNets FIRE NASH!!!.. WTF!' | -0.87 | 0.0 | 0.815 |

**Team:** ***Los Angeles Lakers***

| User ID | Tweet | Compound | Positive | Negative |
|---|---|---|---|---|
| **1520170549137543168** | '@imgregorous @Lakers Haha, amazing' | 0.7783 | 0.773 | 0.0 |
| **1520136187238789120** | 'Man hell no @Lakers https://t.co/cL0cAQ34ln' | -0.7783 | 0.0 | 0.694 |

Sorting and displaying the most positive and most negative tweets help us to determine the reliability and correctness of the sentiment analysis provided by Vader. Throughout most data entries of the Tweet column with the Compound Column, the analysis reflected certain accuracy. However, some issues exist. For example, the most positive and most negative tweet for Trae Young is not correct. “PLEASE WIN” and “No pressure” are closer to neutral sentences rather than being positive or negative. Therefore, Vader's sentiment analyzer has some limitations, and the reasons are:

- **Tokenization**: sentiment analyzer transforms the sentence into separated pieces and performs analysis on those pieces. This process is efficient in terms of processing. However, tokenization cannot identify certain tones.
- **Capitalization**: sentiment analyzer will assign a greater sentiment score for words that have their literal in upper cases compared to those in lower cases.

In this case, "win" will be considered a positive word, and capitalized words will be assigned with a higher score due to a stronger emotion. For the same reason, even though "No pressure" should be considered neutral, "No" itself contains a negative sentiment score. That is why the whole sentence will be evaluated as negative.

```
[In [12]: from nltk.corpus import sentiwordnet as swn

[In [13]: win = swn.senti_synset('WIN.n.01')

[In [14]: print(win)
 <win.n.01: PosScore=0.125 NegScore=0.0>

[In [15]: no = swn.senti_synset('No.n.01')

[In [16]: print(no)
 <no.n.01: PosScore=0.0 NegScore=0.25>
```

The following table shows the general data generated by performing sentiment analysis and statistical analysis on the fetched tweets.

**Compound Average** is the average of compound score sum of all tweets.

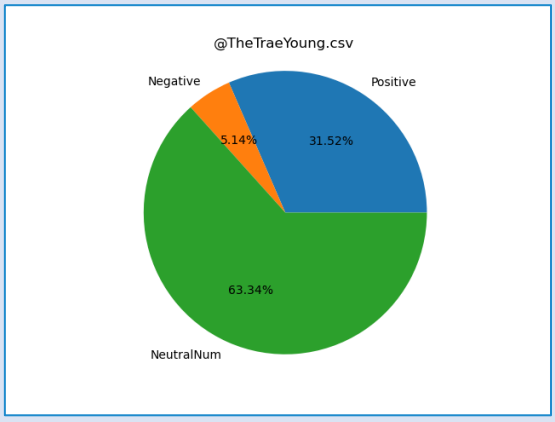
**# Positive** is the total number of positive tweets.

**# Negative** is the total number of negative tweets.

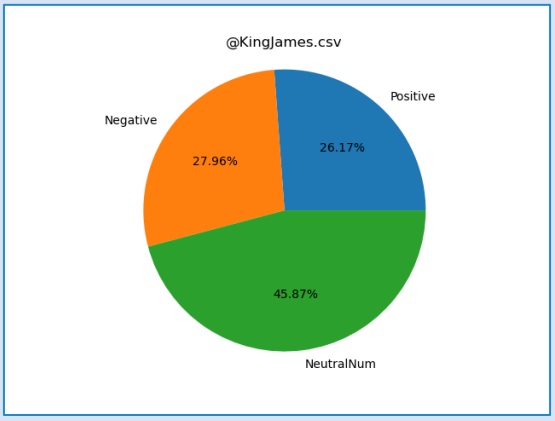
**# Neutral** is the total number of neutral tweets.

**Total** is the total number of tweets

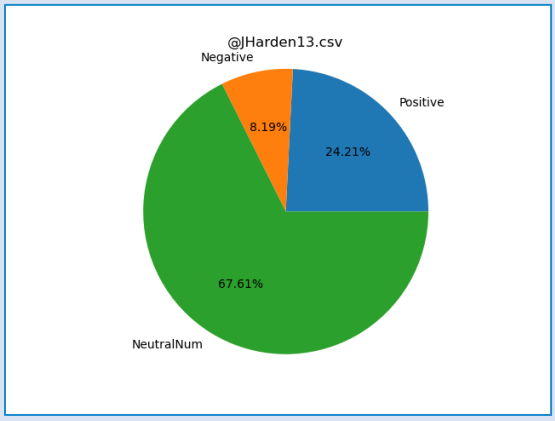
| Name | Compound Average | # Positive | # Negative | # Neutral | Total |
|---|---|---|---|---|---|
| **Trae Young** | 0.1367 | 3048 | 497 | 6126 | 9671 |
| **Lebron James** | -0.0285 | 2247 | 2401 | 3938 | 8586 |
| **James Harden** | 0.0810 | 1576 | 533 | 4402 | 6511 |
| **Atlanta Hawks** | 0.1245 | 1889 | 721 | 2910 | 5520 |
| **Brooklyn Nets** | 0.0474 | 1948 | 1310 | 4131 | 7389 |
| **Los Angeles Lakers** | 0.1106 | 2043 | 869 | 3088 | 6000 |



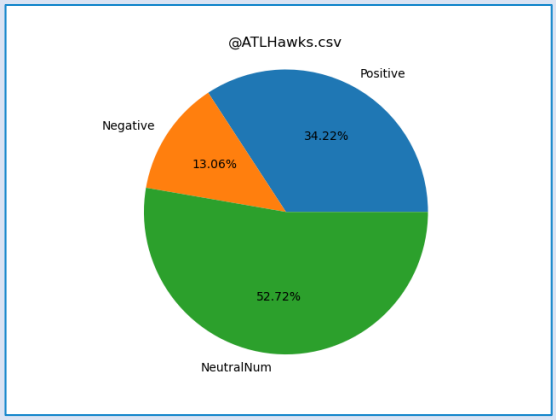

Trae Young has an overall positive reputation among Twitter users. The compound average is 0.1367 and has 31.5% of positive comments, 5.14% of negative comments.



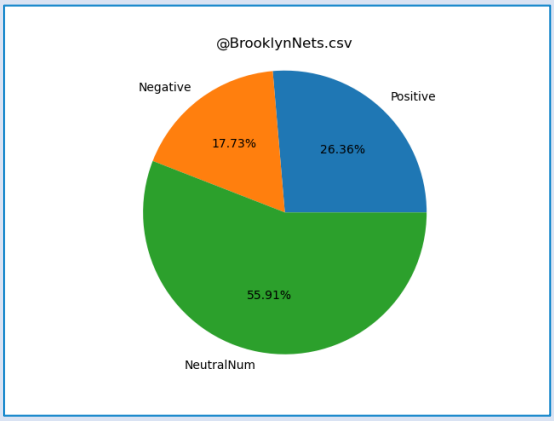

Lebron James has a higher percentage of negative Tweets with a negative compound average of -0.0285, reflecting that Twitter users' attitude toward Lebron James is minorly negative. Lebron James' percentage of negative tweets is 27.96%, greater than that of positive, which is 26.17%.



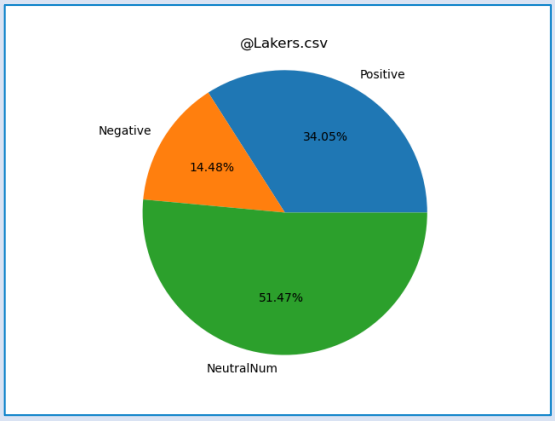

James Harden's compound average is 0.0810. The percentage of positive tweets is 24.21%, which is much greater than that of negative's 8.19%. Similar to Trae Young, James Harden has an overall positive reputation among Twitter users.




The Atlanta Hawks is a team that has an overall positive reputation by analyzing the tweets. The compound average is 0.1245. 34.22% of the tweets reflect a positive attitude, while 13.06% are negative.




The Brooklyn Nets' reputation among Twitter users is closer to minorly positive (0.0474 compound average). Even though there are more positive tweets (26.36%) than negative tweets (17.73%), the positive tweets do not have a high sentiment score, which cannot increase the compound average.

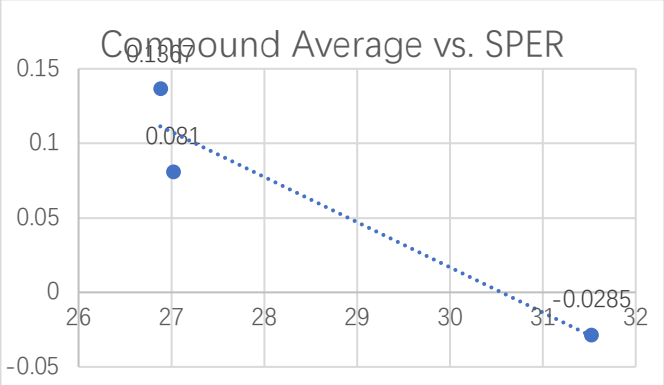



The compound average of the Los Angeles Lakers is 0.1106, which reflects an overall positive tweets attitude. There are 34.05% of total tweets contributed positively, and 14.48% contribute negatively.

**Data Comparison**

**Players**

| Player Name | Compound Average | SPER |
|---|---|---|
| **Trae Young** | 0.1367 | 26.88 |
| **Lebron James** | -0.0285 | 31.52 |
| **James Harden** | 0.0810 | 27.02 |

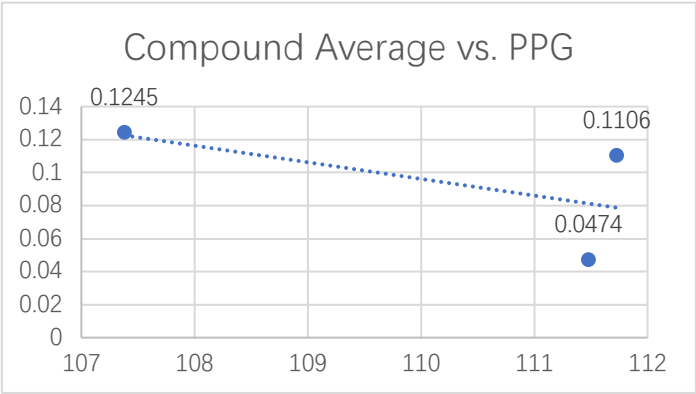



The table above shows the correlation between the player's compound average and SPER. For the player with the highest SPER (Lebron James), his compound average of tweets shows up to be the lowest. Vice versa, for the player with lower SPER (Trae Young), his compound average shows up to be the highest.

The comparison reflects a negative linear correlation between compound average and SPER. In other words, the better the player performs (higher SPER), the lower the reputation of that player is on Twitter (compound average). This contradicts our assumption, which was expecting a positive liner correlation. The reason that causes this contradiction will be presented in the **Additional Analysis** Section.

**Teams**

| Team Name | Compound Average | PPG |
|---|---|---|
| **Atlanta Hawks** | 0.1245 | 107.38 |
| **Brooklyn Nets** | 0.0474 | 111.48 |
| **Los Angeles Lakers** | 0.1106 | 111.73 |




The table shows the correlation between the teams' compound average and PPG. Similar to the comparison for players, the lowest PPG is matched to the highest compound average, while the teams with better performance are assigned with a lower compound average.

This also contradicts our prediction, we are expecting a positive linear correlation between compound average and PPG, but the negative relation is given. The reason that causes this situation will also be included in the **Additional Analysis** Section.

# VII. Additional Analysis

**Does graphics satisfy predictions?**

In our original conjecture, we presumed that a player with a higher SPER (Simpler Player Efficiency Rating, a value that authenticates a player's performance) or a team with a higher PPG (Points Per Game) is more likely to have associations with tweets and comments with a higher compound average (The sum of positive, negative, and neutral manner scores which are then normalized between -1(most extreme negative) and +1 (most extreme positive) among comments on Twitter). In other words, our paper was made under the assumption that a player's performances in a year should draw a positive linear relationship with a more positive sentiment analysis feedback, which was also a belief that was trusted by the public – winning more results in better evaluations among a player or team.

However, in fact, it was quite the opposite. Whether among players or teams, the scale of actual performances to evaluations was desperately different. The universal correlation graphics demonstrated that such public belief is unwarranted. The actual relationship between SPER/PPG and Compound average is more likely to be contradictory.

**Why is this happening?**

In a player's case, from the overall situation we conjecture that rather than exalting a player based on his individual statistics, the public manner attended greater interest in a player's contributions to the whole team.

To figure out its mystery, NBA players' status quo and associated factors have been taken into consideration. For instance, although James Harden had the second-highest SPER value among the three players chosen, many commentators impugned him for he was just recently bought by team Sixer but did not get his team into the playoffs. Tweets asserting his irresponsibility and apathetic performances, claiming that he needs changes.

One contradiction to James Harden's public reflection is commenting on Trae Young, he had the lowest SPER value, yet had an extremely positive compound average compared to the other two players. Commentators considered Trae Young as a young player who joined the league for four seasons so far yet exhibiting predominant potential on points gaining and assistance to teammates. The public intends to demonstrate greater empathy for inexperienced players rather than experienced players.

As a further bolster to the above statement, Lebron James, as both the youngest and the oldest player who claimed an average of 30+ points per game in the history of the NBA and one

of the best basketball players who played over 19 seasons, received a -0.0285 compound average for his latterly miss on not properly leading the Los Angeles Lakers while being the bellwether of his team. Comments on Twitter describe Lebron as a player recently trying to get better scores rather than guiding his teammates. Furthermore, comments claimed that he is dubious for actions like giving up team defenses so that his personal rating would not be affected by any defense score, which is an inappropriate action that is hardly tolerated by the public.

In the team's case, although team Atlanta Hawks and Los Angeles Lakers' PPG value did somehow conduct a trivial negative scale with their compound average, their difference is petty and thus are concluded as acceptable biases. However, in spite of the fact that team Brooklyn Nets received a compound average of 0.0474, which is 0.0771 less than the Atlanta Hawks' compound average value and 0.0632 less than the Los Angeles Lakers while generating a paramount PPG, it is reasonable to draw a conclusion that a team's PPG values don't utterly reflect its overall evaluation from the public.

# VIII. Conclusion

This paper conducts an overall analysis of NBA players and teams based on authentic statistics from the official NBA league and tweet comment reflections from Twitter. Throughout the paper's constructions, multiple tools including APIs, Sentiment Analysis, and Python graphic tools have been used to parse essential data including all players/teams authenticated statistics from season 2000-current, their correlated graphics, and over 44000 tweets with sentiment analysis calculation on each one of them.

From all data we collected, we concluded that based on statistics Lebron James performed the best of the three players we chose in season 2021 with a SPER of 31.52, followed by James Harden with 27.02, and finally Trae Young with 26.88. Considering player's promotion by season 2020, Lebron once again claimed top place with an advance of 13.59% on his SPER, followed by the new start in NBA Trae Young with an advance of 9.31%, and a 13.78% deterioration on James Harden. Los Angeles Lakers perform the best among all three players with a PPG of 111.73, followed by Brooklyn Nets with 111.48, and finally Atlanta Hawks with 107.38, each performing roughly the same.

However, comments from the public draw a quite opposite outcome. From the sentiment analysis' result we know that Trae Young received the overall most positive feedback with a compound percentage of 0.1367, followed by James Harden with 0.0810, and finally Lebron James's -0.0285 average. For teams' evaluations, Atlanta Hawks won the comparison with an 0.1245 compound average, closely followed by Los Angeles Lakers with 0.1106, then finally Brooklyn Nets with 0.0474. Both Lebron James and Brooklyn Nets receive mysteriously low compound average while they were performing as the top tire within comparisons to the other two players/teams.

The two sets of data we collected seem to build disparage. We can conclude from all the above completion that besides a player/team's statistics data, rather than exalting a player based on his individual statistics, the public manner attended greater interest in a player's contributions to the whole team. Which is supported by the fact the comments among a player or team are jointly related to the player/team's attitudes to games. While inexperienced but passionate Trae generated a lesser SPER but received more encouragement and compliments, Lebron and his Los Angeles Lakers received mounts of critiques for no longer treating the games seriously even though his stats may look good.

# Reference


Balldontlie. (2022). Introduction. *Balldontlie.* Retrieved May 7th, 2022, from:

https://www.balldontlie.io/#introduction

Benson, P. (2022). Atlanta Hawks Evaluations: Trae Young. *FanNation*. Retrieved May 9th, 2022, from:

https://www.si.com/nba/hawks/news/atlanta-hawks-evaluations-trae-young#gid=ci02a052bca0002793&pid=offense---a

Beri, A. (2020). SENTIMENTAL ANALYSIS USING VADER. *Towards Data Science*. Retrieved May 7th, 2022, from:

https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

FinancesOnline. (2022). Number of Twitter Users 2022/2023: Demographics, Breakdowns & Predictions. *FinancesOnline*. Retrieved May 8th, 2022, from:

https://financesonline.com/number-of-twitter-users/#:~:text=How%20many%20Twitter%20users%20are,a%2024%25%20growth%20from%202019.

NLTK. (2022). Documentation. *NLTK*. Retrieved May 7th, 2022, from:

https://www.nltk.org/

Tweepy. (2022). Documentation. *Tweepy*. Retrieved May 7th, 2022, from:

https://docs.tweepy.org/en/v3.5.0/index.html