

Report

Data manipulation

The TweetsNBA dataset has been cleaned in order to make it easier to analyse the data. Firstly, I decided to remove unnecessary columns for the analysis based on two purposes: to reduce the size of the whole storage and exclude the fields that will not give any conclusive results due to a large number of NaNs. TweetsNBA dataset has several columns with IDs in two types of variables (integer and string) thus in order to reduce the size of the storage, all the columns with IDs in integer type have been dropped. Columns "geo", "coordinates", "place" have been deleted due to a large number of NaNs. This decision was also made because the user data has information about their location and it is mostly filled in. The remaining columns were dropped due to the fact that there are many empty fields. The list of columns that have been removed is:

- 1.possibly_sensitive
- 2.quoted_status_permalink
- 3.in_reply_to_status_id
- 4.in_reply_to_user_id
- 5.quoted_status_id
- 6.display_text_range
- 7.geo
- 8.coordinates
- 9.place
- 10.contributors
- 11.id

Secondly, I decided to add a popularity attribute to an existing collection based on the sum of quote_count, reply_count, favorite_count and retweet_count. Using the aggregation framework new fields have been created for the purpose of further analysis on what is the top tweet in the dataset. Important note that all actions carry equal weight in a query.

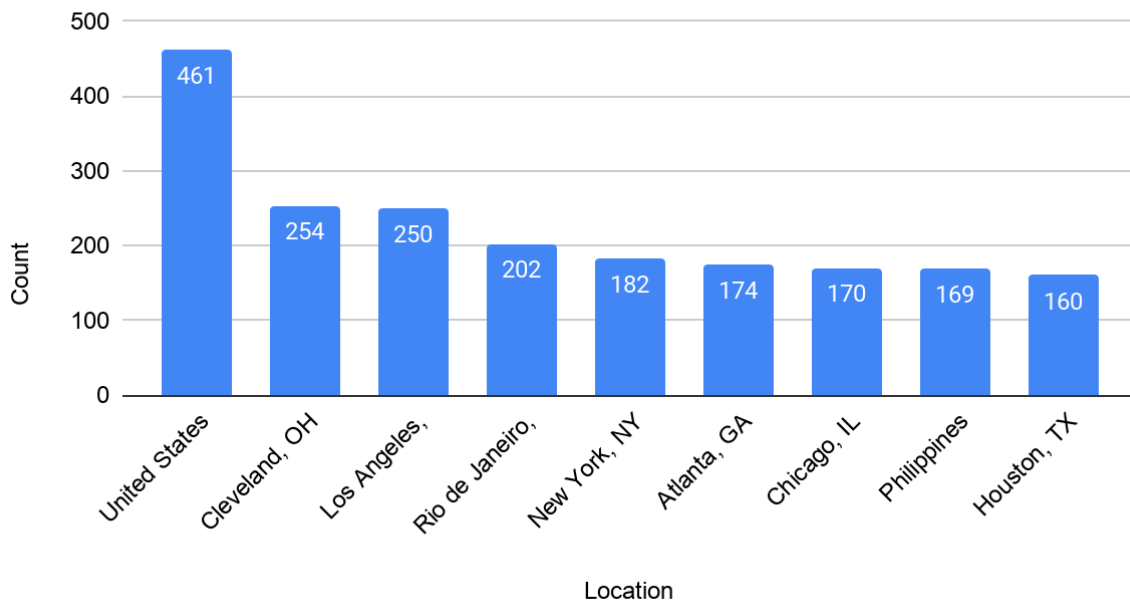
1. Top 10 trending hashtags

I construct a query that produces the top 10 hashtags along with how frequently they occur in the NBA finals tweets data. Looking at the result, the #NBAFinals was used the most with a count of 40,120. Followed, was #WhateverItTakes with a count of 13,687. The third most frequent hashtag was found to be #DubNation which is the hashtag associated with GoldenState Warriors with a count of 1,909 and this is more than double of the hashtag associated with Cleveland Cavaliers as #Cavs amounted to only 808. This could possibly mean there were more people rooting for golden state warriors to win the finals.

2. The most popular place from where the tweet was placed/ which and how many places

Here, I construct a query that produces the location where tweets were posted from and how many tweets came from that area. We can see from the results that 14,484 tweets are coming from a null location. This could suggest that the users have their location services off, possibly for privacy reasons. Unsurprisingly, the United States and cities like Cleveland, Los Angeles, New York, Atlanta and Chicago were part of the top 10 locations where tweets were posted from. On the other hand, what was surprising about the results was the occurrence of Brazil and the Philippines have a count of 202 and 169 respectively. The low counts in the top places shown below suggest that there are no real epicentres of NBA fans watching the third game. The counts are very much spread out, assuming that users with 'null' location are not heavily concentrated and on average follow the same pattern as the rest of the data.

Tweet count by location



3. Followers count with verified and unverified users

In this query, the aim was to discover the number of followers both verified and unverified users have. The top 3 unverified users with the most followers were SuoerElmo with 1,279,999 followers, DZMMTeleRadyo with 1,279,317 and Javisness with 1,066,235. These people are influential and could be targeted by certain businesses if they want to reach a wider audience. On the other hand, the top 3 verified users include NBA, the official twitter account of the NBA, with 27,814,570, Marcos Mion who is a Brazilian TV actor, with 13,498,273 which could be the reason for the high occurrence of viewership in Brazil. Finally Bleacher Report, a sports page with 6,660,591.

4. Top languages used for the tweets

I constructed a query that returns the different types of languages used across the tweets data along with the number of users using each one. From the result, we see that English(en) is the most used language with a count of 37,756, which doesn't come as a surprise as the United States was top for prime tweet location. Spanish(es) was next with a count of 3,164. Arabic(ar) was also widely used as it had a count of 2,129. Portuguese(pt) was also widely used with a count of 1,901, this could be a result of the large viewership from Brazil since Portuguese is their official language. Lastly, we have Turkish(tr) with a count of 1,205. From this, we can see the diversity of languages across tweets and it's interesting to know the number of people tweeting in these different languages.

Top languages used for tweets

Tagalog - Filipino

1.2%

Turkish

2.4%

Portugese

3.7%

Arabic

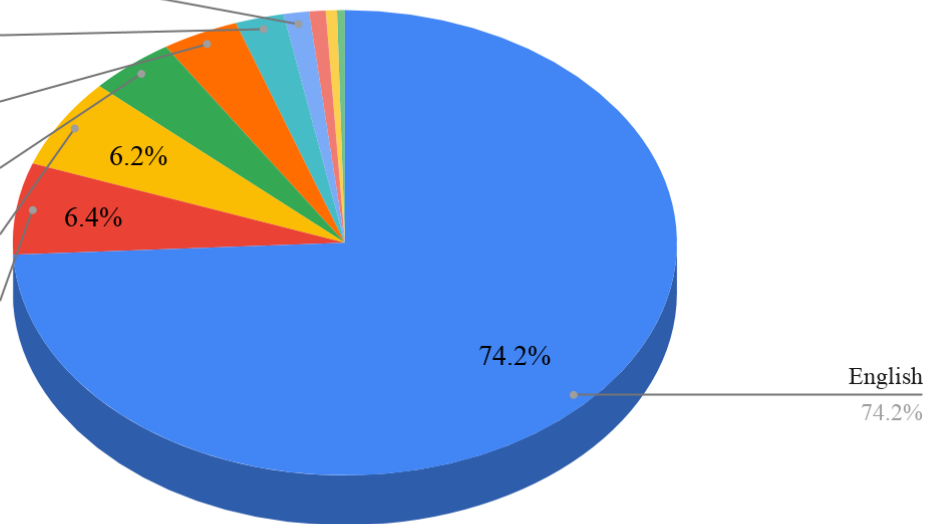
4.2%

Espanol

6.2%

Undefined

6.4%



5. Most used source

The aim of this query was to find out what platform users were tweeting from and how many users were using these different platforms. Twitter for iPhone was the most used platform with a count of 27,397 followed by twitter for android with a count of 16,586. This shows that more people across the data use Apple products rather than android. Twitter lite which is an app for android was used by some users with a count of 1,173. Tweetdeck which is an app for laptops was also used as this amounted to 706 users tweeting from this app. Here are the top 5 sources.

Most used source

Tweetdeck

1.4%

Twitter lite

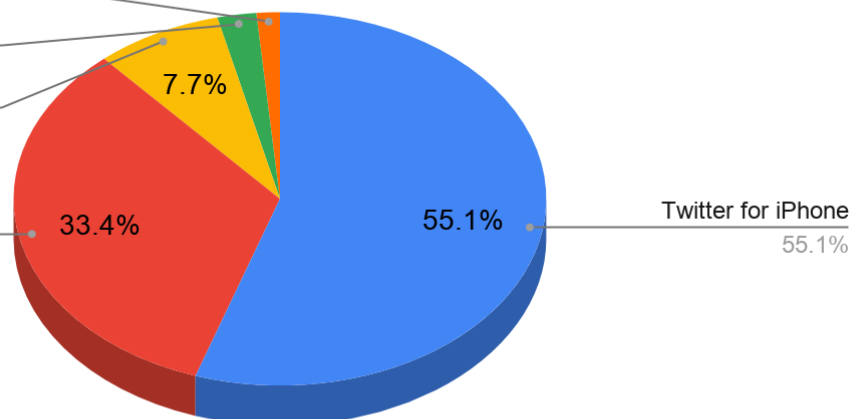
2.4%

Twitter web client

7.7%

Twitter for Android

33.4%



6. Most retweeted tweet

In this query, I discover the most retweeted tweet, the number of retweets and the user who posted the tweet. The user with the most retweeted tweet was the NBA's official twitter account with 10,016 retweets. The tweet was "🏀🏆 OH YES HE DID! 🏀🏆\n\n#WhateverItTakes #NBAFinals along with a video and this was in reference to a play LeBron James made to earn his team 2 points. The 2nd most retweeted tweet was the Bleacher report tweet with 4,076 retweets for a tweet about LeBron James dunking the ball and players around him looked in awe. The 3rd most retweeted tweet was Marcellus Wiley's tweet with 3,608 retweets. People like tweeting about live events and verified accounts like the NBA and Bleacher report are reputable sources so people tend to retweet their tweets a lot more than unverified accounts. These accounts have a much larger following so tweets posted are seen by a lot of viewers which can explain the high retweet count that they both have.

User	Text	Count
NBA	"🏀🏆 OH YES HE DID! 🏀🏆\n\n#WhateverItTakes #NBAFinals https://t.co/Axun0gOmok "	10,016
Bleacher Report	"Shook.\n\n#NBAFinals https://t.co/9JVTJcPEl0 "	4,076
Marcellus Wiley	"I'm 1000% sure that LeBron can't win this series! Soooo, if the #Cavs beat the #Warriors in the NBA Finals, I'll g... https://t.co/h5zlgazRab "	3,608
FirstTeam	"JEU CONCOURS #NBAFinals \nGAGNES LES SHOES DE TON JOUEURS PRÉFÉRÉ\n\nFollow @FirstTeam101 & @parisbasketball \n+ RT ce... https://t.co/wUDZh3ZR3N "	774
NBAonTNT	"OH MY, LEBRON JAMES! 🤔\n\n#NBAFinals \n\n https://t.co/ZoeHxiMFee "	749



7. Which team has more twitter user fans

The amount of times Cavs were mentioned in a tweet amounted to 1,571 and the number of time Warriors was mentioned in a tweet totalled 2,528. This could suggest that GoldenState Warriors had more people on twitter rooting for them. This can also be seen when we analysed the number of hashtags used as #DubNation had a lot more than #Cavs.

8. The top tweet

The aim of this query was to find the tweet with the most interactions. That includes favourite count, retweets, quote count and replies. Here are the top 5 tweets

User	Text	Popularity
Rico	"Stop being scary bro she right there go talk to her 🤔 https://t.co/JDurREDxj "	131,616
UpTheThunder	"Snakes on a Plane (2006) https://t.co/1lhHbMjU3o "	32,811

NBA	"  OH YES HE DID!  \n\n#WhateverItTakes #NBAFinals https://t.co/Axun0gOmok "	28,548
NBA	"The BEST of Steph Curry (31 ppg, 6.5 rpg, 8.5 apg) from the @warriors W's in Games 1 & 2 of the #NBAFinals!... https://t.co/sZdSfRL4qq "	19,301
NBA	"Stephen Curry (33 PTS, 8 AST, 7 REB) knocked down an #NBAFinals record 9 3-pointers to fuel the @warriors Game 2 wi... https://t.co/EdxG5IG9ms "	18,385

The top tweet was a meme of Lebron James looking confused at JR Smith after he ran down the clock thinking they were ahead in the game and unfortunately led to the Cavs losing the game. The second tweet, however, has nothing to do with the NBA. rather, it was a tweet with a picture of Hilary Clinton being referred to as a snake. It's interesting to see that the top two tweets were funny tweets and had little or no correlation with the game that was going on. This could suggest that users appreciate the humour in tweets even at a time when the NBAFinals is going on.