Data Analysis of NBA Twitter - Anpandoh - Medium

By Anpandoh

Source: https://medium.com/@anpandoh/data-analysis-of-nba-twitter-8265e95e5a2

With so much concern over data privacy over the last few years Zachary Youngblood → https://medium.com/u/8d50ecea0d3c?source=post_page-----8265e95e5a2---------- and I decided to dive into our own data given the personal data transparency that Twitter has provided. We are both fairly active Twitter users and wanted to put what we learned in our Computational Analysis of Big Data. We both hit the download button, got our data, and dissected the results. An important thing to keep in mind is that Zach and I had different use cases for Twitter as I mostly focused on NBA and Basketball related Twitter whereas Zach had a broader topic list more concerned with business and technology.

Initial Predictions

You will be seeing a lot of data related to Damian Lillard and the Portland Trail Blazers, so to get you familiar I asked chatGPT to give you a quick summary of those topics:

> Damian Lillard is a professional basketball player who has spent his entire career with the Portland Trail Blazers. He is a six-time NBA All-Star and one of the best point guards in the league. Lillard is known for his clutch performances in important games and has earned the nickname "Dame Time" for his ability to hit game-winning shots. He has also been a leader on and off the court for the Blazers and is widely regarded as one of the most beloved and important figures in Portland sports history. Despite the team's struggles in recent years, Lillard's loyalty and commitment to the city and the Blazers organization have made him a fan favorite and a symbol of resilience for the city of Portland.

I predicted to also see a lot of negative things about Terry Stotts, the former Trail Blazer coach who I was not a fan of. Although he got fired a couple of years ago, it was also the period when I was the most active on Twitter. Besides that, we will just have to dive into the data and draw some conclusions from the numbers.

Collecting and Sorting the Data

Once we had the data download into JSONs we extracted all of our liked tweets and Twitter's inferred interests relating to us. We then had to sort these likes tweets into these interests categories which is why we used semantic sentence similarity. We did this by using the Huggingface model: all-MiniLM-L6-v2 → https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. After multiple hours we were able to get normalized similarity values between each tweet and interest category:

	Stock AAPL	Stock AMD	Stock AMZN	Stock BA	Stock BABA	Stock BAC	Stock BYND	Stock C	Stock CCL	Stock CRM	...	USA	University of Alabama
Caitlin Clark will heal the nation	0.036086	-0.070474	-0.012616	-0.010784	-0.020514	0.025537	-0.018438	0.030171	0.064532	-0.101122	...	0.038209	0.110159
Do your workkkk @ChuckyHepburn 🙏🙏🙏🙏🙏 https://t.co/BZI4OwVTO1	0.039882	0.031351	0.117099	0.109217	0.114864	0.113702	0.085848	0.098703	0.101533	0.026367	...	0.050312	0.071210
This hesi from Caitlin Clark was tough 🔥 \n\n#WFinalFour https://t.co/QFL4rssOdc	0.126992	-0.038261	0.086859	0.117749	0.122893	0.119531	0.096982	0.126597	0.133412	0.002160	...	0.072238	0.169980
@jeffborzello @hoopz_wisconsin @ChuckyHepburn get him to @BadgerMBB	0.068763	-0.006356	0.082374	0.107663	0.035654	0.099572	0.043533	0.037310	-0.009315	0.023671	...	0.075760	0.210383
https://t.co/oa5HsKvmOS	0.029574	0.046315	0.088244	0.092580	0.071733	0.094284	0.028250	0.021087	0.074702	0.003115	...	0.179241	0.112203

Snippet of Similarity

From there we took the top 3 most similar categories for each tweet and removed tweets that did not have similarity values above 0.4. Then we transformed the dataframe to have a 1 for an interest category that the tweet belongs to and 0 otherwise:

]:

	Stock AAPL	Stock AMD	Stock AMZN	Stock BA	Stock BABA	Stock BAC	Stock BYND	Stock C	Stock CCL	Stock CRM	...	Wisconsin Badgers
@jeffborzello @hoopz_wisconsin @ChuckyHepburn get him to @BadgerMBB	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0
Josh Giddey, Tyrese Maxey, Mobley, DG, A Simmons, JJJ, Naz Reid, Trey Murphy, Cam Thomas, Jalen Williams (OKC), KPJ, Tyler Herro, Jordan Poole, Bane, Clax, WCJ, Shaedon Sharpe, Sengun, J Allen, MPJ, RJ Barrett, Mathurin, Jaden McD, Kevin Huerter. its more but only 140 characters https://t.co/KXqB67uMbt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
Shaedon Sharpe 😮😮😮 \n\nFlying blocks >> Flying dunks ????	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0

Snippet of Categorization

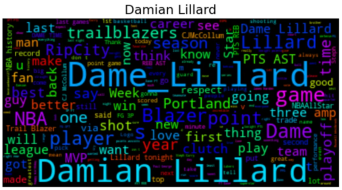
Word Clouds

One easiest and nicest ways to visualize our data is generating word clouds for different interest topics. First after having categorized all the tweets to the different interest topics we list the interest topics with the most tweets associated with them by summing up the columns (interests) which unsurprisingly has Damian Lillard as number one. We can generate the word cloud by combing all the tweets into one big text string and feeding it into the wordcloud library. It is also important to note that

we filtered out some words such as “https” for links, and “suspended” for tweets that are not shown because the account has been removed. The wordcloud for all the tweets is as follows, with individual interests categories listed next:

Overall Word Cloud

Damian Lillard



We can see words like “fan, best, love, good, MVP, and star” that properly depict my feelings toward the topic.

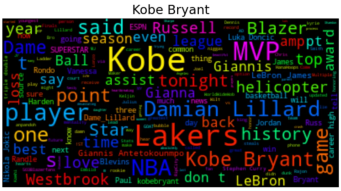
NBA stats

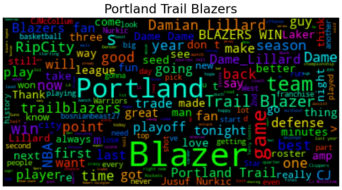
Here we can see all the important NBA stats listed including: “PPG, RPG, APG, Assists, FG, 3PT, clutch time, blocks,” etc.

NBA Officials

I thought this one was intriguing as well as it includes players like Harden, Rondo, and Lebron who are all known for referee interaction whether it be drawing fouls or complaining with refs.

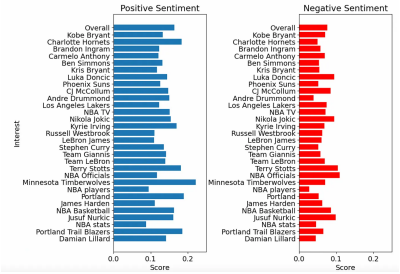
Some more interesting ones:





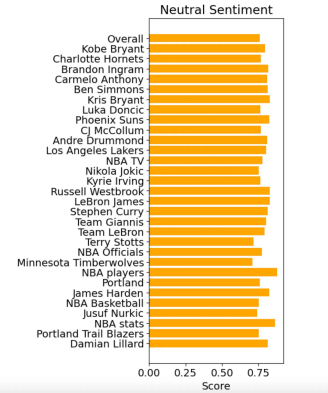
Sentiment Analysis

After generating a word cloud, we conducted a sentiment analysis of the top 30 interests and overall sentiment as a means of establishing a baseline. The findings of this analysis proved to be the most intriguing aspect of our project, as many unexpected results emerged.



In particular, while the high positivity analysis of tweets concerning Portland did not come as a surprise, what was disconcerting was discovering that Damian Lillard, my favorite player, had a lower sentiment score than the overall sentiment. Equally surprising was the fact that the Minnesota Timberwolves, a team I had previously considered to be unremarkable, had the highest sentiment score by a signifi-

cant margin. This could be important to keep in mind should the Blazers ever leave Portland. Another notable discovery was that Terry Stotts, a coach whom I believed to be inadequate and advocated for his firing (which subsequently occurred), had a relatively high sentiment score, however, the negativity graph gave more reasonable findings as he had the second-most negative tweets related to him, surpassed only by those critical of NBA referees, which says a lot. In my opinion, the negativity graph provided the most accurate representation of my general NBA opinions, with some of the lowest negative sentiment scores attributed to Damian Lillard, NBA players, and the Portland Trail Blazers, which is unsurprising. The only exception was Jusuf Nurkic, who, despite having been a personal favorite of mine historically (maybe not recently), had a comparatively low score.



While the neutrality data was relatively less useful than the other two graphs, it has its significance, particularly when examining the lower scored bars which indicate more controversial interests such as Kyrie Irving, who has been a subject of controversy throughout his career, and Terry Stotts.

Bag of Words and PCA Analysis

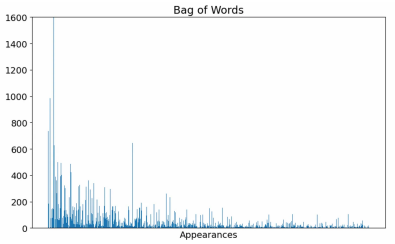
The next step for us is to collect how many times a word showed up for each interest by creating a bag of words matrix. This allowed us to do things such as measuring and visualizing word frequency and fitting a PCA on the words to explain variance. We created this matrix by first including a function to clean the sentences into a list of words and then looping through them for all the tweets in each interest. From this, we were able to create the following matrix:

	damelillard	man	benmclemore	yall	know	guy	rock	chalk	thekidet	frog	...	crowd	adversity	continue	pregame
Damian Lillard	442	27		1	6	23	13	3	1	1	1 ...	2	1	1	1
Portland Trail Blazers	33	22		1	1	18	11	0	0	0	0 ...	0	1	1	0
NBA stats	29	9		0	1	5	5	0	0	0	0 ...	0	0	0	0
Jusuf Nurkic	1	7		0	1	5	6	0	0	0	0 ...	0	0	1	0
NBA Basketball	8	7		0	4	13	0	1	0	0	0 ...	0	0	1	0
...	...	...		...	...	...	...	...	...	...	... ...	...	...	...	...
Keenan Allen	0	0		0	0	2	2	0	0	0	0 ...	0	0	0	0
Charles Barkley	5	1		0	0	0	0	0	0	0	0 ...	0	0	0	0
Rodney Hood	0	2		0	0	0	1	0	0	0	0 ...	0	0	0	0
College Football	0	0		0	0	0	1	0	0	0	0 ...	0	0	0	0
College basketball	1	0		0	0	2	0	0	0	0	0 ...	0	0	0	0

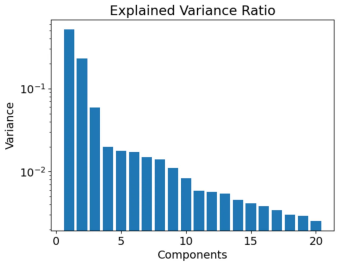
100 rows × 2863 columns

```
Top 10 most used:
blazers          1607
lillard          1268
nba              1092
damian            984
game              957
portland          945
dame              880
damelillard       737
team              646
season            627
one               515
points            501
games             494
win               485
pts               451
like              448
```

Visualizing this we can see the sheer quantity of times I have referenced the Blazers and words related to Damian Lillard. We can see the major spikes of data come near the beginning, due to the fact that it is in order of how tweets are fed to the matrix and has nothing to do with the actual data.



After this visualization, we can also perform Principal Component Analysis, where we split the data into 20 components and using some linear algebra trick we can see the explained variance ratio graphed as so:



This basically means that the first 2 components, which capture 75% of the total variance, account for the most significant trends and patterns, where each component is the variance that is completely uncorrelated from each other. This also means the remaining components capture more subtle details that may be less important.

Conclusion

In conclusion, by utilizing various techniques such as word clouds, sentiment analysis, bag of words, and PCA analysis, we were able to gain insights into our data and understand the trends and patterns. We learned that Damian Lillard was the most tweeted about player, but surprisingly had a lower sentiment score than the overall sentiment. The Minnesota Timberwolves had the highest sentiment score, indicating that they are more popular than previously thought. The negativity graph provided the most accurate representation of our general NBA opinions, and the bag of words matrix showed the sheer quantity of times we referenced the Blazers and words related to Damian Lillard. PCA analysis revealed that the first two components account for the most significant trends and patterns, capturing 75% of the total variance, while the remaining components captured more subtle details that may be less important. Overall, these techniques allowed us to extract meaningful insights from our data and gain a better understanding of our interests and opinions related to the NBA.

Our Joint General Reflection

The analysis of Twitter data can be a daunting task, requiring significant time and effort to clean, categorize, and format the data. In our analysis, we found that loading the data from the JSON was relatively straightforward, but categorizing our liked tweets using semantic analysis from a Hugging Face API took multiple hours. However, this step was critical in enabling us to draw conclusions about individual interests and identify patterns within the data.

Despite the challenges of formatting our data correctly, we were able to apply various Natural Language Processing techniques to visualize patterns and trends. The use of word clouds provided an interesting visual representation of the most commonly used words in our data. We found that sentiment analysis provided surprising results, highlighting the varying emotional tones in the tweets we analyzed.

Furthermore, by using a bag of words matrix to count the frequency of words and performing principal component analysis (PCA), we were able to identify variance within the data. This enabled us to understand the distribution of our data and helped to identify any underlying trends or patterns that were not immediately apparent.

The process of analyzing the Twitter data yielded interesting and insightful results. By applying various NLP techniques, we gained a deeper understanding of the content and sentiment of the tweets we analyzed. Additionally, our findings provided us with new insights into the interests and preferences of the users whose tweets we examined that we never imagined in the first place.

Overall, this analysis demonstrated the potential of NLP techniques in uncovering meaningful insights from large datasets. While the process of cleaning and categorizing the data was time-consuming, the results proved to be valuable in understanding the underlying patterns and trends within the Twitter data. This analysis provides a useful example of how NLP techniques can be applied to social media data and highlights the potential for further exploration in this area.

Be sure to check out the code → https://github.com/Anpandoh/AneeshWiscProjects/tree/55f664001f157cce279da1774a0e16b7ec1ec7d3/Big_Data/finalProject. *And here is a link to Zach's article if you want a broader range of data rather than just NBA Basketball Tweets.*