Assessing NBA Players' Political Impact through Twitter

Ben Thorpe

November 21, 2020

Introduction

The social activism of professional athletes in the United States has grown substantially over the past 10 years, and much of this increase in spreading awareness has been possible due to the rise of social media. Players are able to post about issues that are important to them whenever they desire, and with the fight against systemic racism and the polarizing presidential election coming to a head over the past few months, they have done just that. Whether it has been through calling for government action after the unjust killings of George Floyd and Breonna Taylor at the hands of police officers or purposefully reminding their followers to get registered to vote, athletes have been using their platform to promote social justice initiatives extensively over the past few months in particular. However, there are plenty of Americans who disagree with this progression and would like them to just "stick to sports." For example, Fox News anchor Laura Ingraham notably told LeBron James to "shut up and dribble" after he publicly expressed his disdain for President Trump in 2018. Ingraham and other detractors of these acts of player empowerment have called for Americans to boycott the league this past season as they felt sports should be devoid of any outside messages, especially topics deemed to be political. They also commonly claimed that these messages were going unheard and would not affect the mindsets or actions of Americans. With this in mind, I set out to measure the impact athletes have through social justice posts on social media and to analyze how the public interacts with their political tweets. I am focusing on players in the National Basketball Association (NBA), since they have been the most outspoken on social justice issues.

Hypotheses

In order to measure the social justice and political impact of NBA players, I am analyzing their recent Twitter data. Since the focus of this study is on their social activism, I am identifying which tweets are related to this topic for analysis. Since I am analyzing the data in a variety of ways, I have come up with multiple hypotheses relating to my main research topic.

1. If a tweet is political, it will have more retweets than non-political tweets, on average (Molyneux, 2017).
2. Important NBA player statistics (points, assists, rebounds, and minutes played) (Li, 2014) and social media metrics (number of retweets and favorites) (Molyneux, 2017) of a tweet will both have an effect in predicting if a tweet is political.
3. On average, political tweets will be more negative than non-political tweets (Bakliwal, 2013).
4. The linear statistical model will be able to predict how many retweets a tweet receives accurately, and the logistic statistical model will be able to determine whether a tweet is political accurately.

Data Wrangling

I first gathered the player statistics and Twitter handles of all NBA players through datasets on the website Basketball Reference and joined them together so each row of data consisted of a player's season stats along with their twitter handle. I then wrangled the last 30 tweets of all NBA players who played in more than five games in the 2019-2020 season using the Twitter screen names from the first dataset. Players who played fewer than six games were removed in order to filter out players who were not on an NBA roster throughout most of the season. However, I did not set a minimum number of minutes played since I wanted to analyze how NBA players tweeted as a whole, which includes the athletes who rarely play but remain rostered. In total, 12,740 tweets were gathered from 449 NBA player (this is less than 30 per player since some athletes have tweeted less than 30 times since their account was started). The player stats and tweets datasets were then merged together by each athlete's respective Twitter handle so that each row in the new dataset included a player's full season-long basketball stats as well as one of their tweets and the accompanying metadata. The full list of variables analyzed is shown below.

```
##  [1] "screen_name"     "Player"          "G"               "MP"
##  [5] "TRB"             "AST"             "PTS"             "text"
##  [9] "favorite_count"  "retweet_count"   "quoted_text"     "retweet_text"
## [13] "followers_count" "political"
```

Where `G` is games played, `MP` is total minutes played, `TRB` is total rebounds, `AST` is total assists, and `PTS` is total points.

I then created a subset of this dataset which only included tweets which were deemed to be political or social justice related. I classified the tweets using a dictionary based approach, where if any tweet, retweet, or quote tweet an athlete posted contained a word from the dictionary its `political` value would be equal to one, and if it did not `political` would be equal to zero. I then filtered for rows where `political` was one to make the `political_tweets` dataset. I manually created the dictionary by including political words that I thought of on my own as well as inspecting the Twitter timelines of NBA players known for being outspoken on social issues and finding language they used often which related to this topic. The dictionary contained a total of 43 words, which are listed below.

```
## [1] "black, democracy, justice, breonna, taylor, trump, Biden, blm, corona, equality, racism, floyd,
```

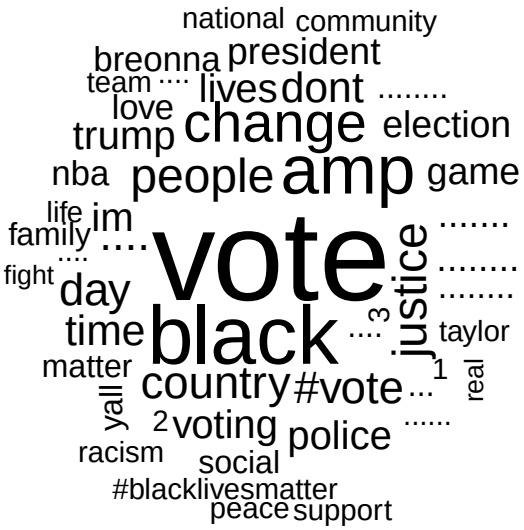
Here are the top 10 athletes regarding how many "political" tweets they have made out of their last 30 tweets posted.

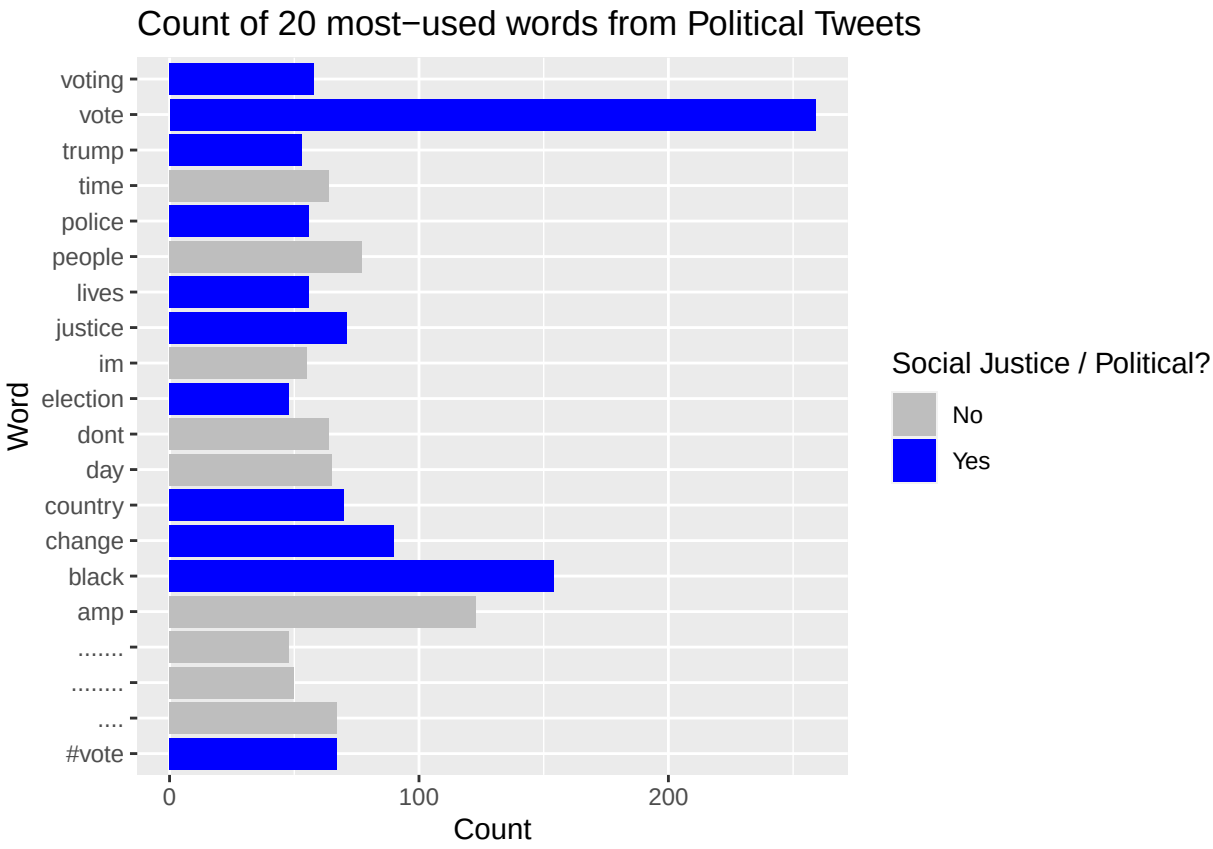
Player	Tweets
Harrison Barnes	21
Damion Lee	20
Sterling Brown	20
Jake Layman	18
Jaylen Brown	18
LeBron James	17
Marquese Chriss	16
Rondae Hollis-Jefferson	14
Shake Milton	14
Chris Paul	13

The list demonstrates that the dictionary approach works at least moderately well since multiple high-profile players who have been the most outspoken on social justice issues, such as LeBron James, Jaylen Brown, and Chris Paul, are on it.

To additionally test whether my dictionary-based approach worked, I wanted to find the most popular words found in the tweet text of tweets classified as "political." I used the `stringr` package

to gather all of the individual words from the tweets and remove "stop words" (words that will always show up often such as "and" and "the"). Below is a word cloud showing the most popular words and a bar graph of the 20 most common words seen in the political tweets dataset.





Methodology

To find which variables were most predictive of retweets, I created a linear model with points, assists, rebounds, minutes played, whether a tweet was political, the number of followers the "tweeter" has, the number of favorites the tweet has, and the interaction between the follower count and favorite count as the independent variables. I chose the retweet count to be the response variable because it is the metric that is most indicative of how many people read the content of a tweet. The more often a tweet is retweeted, the more people the message it contains will reach. Thus, I felt this was the best way to measure the impact of a political tweet. The first four variables are all season totals from the 2019-2020 NBA season, and I included them because these are the first statistics people think of in determining how good a basketball player is. Higher values should lead to more national exposure and thus more interaction with tweets by those athletes. I also wanted to account for the number of followers the user had and the number of favorites a tweet received since I expected these to have a very strong effect on the number of retweets. The interaction variable between these two metrics was included because they should be heavily related, and I wanted to account for this in the model.

In determining which factors were most important in the classification of a tweet as political, I decided to test a logistic model. It had a very similar structure to the linear model described above, except now the `political` variable is the response variable and the favorite count is a factor. The rest of the independent variables are the same as in the linear model. I decided to use a logistic regression in this case since the response variable is a binary categorical variable with the categories being political and not political. Furthermore, by running a logistic model I was able to control for variables that I believed would affect the political nature of a post: a user's important basketball season totals and their social media metrics.

My next test was to figure out whether the political tweets were more negative than non-political tweets

term	estimate	p.value
(Intercept)	3.64e+03	0.000
PTS	5.05e-01	0.673
AST	2.22e+00	0.455
TRB	2.00e+00	0.299
MP	-2.23e+00	0.003
political	7.03e+03	0.000
followers_count	-3.48e-04	0.000
favorite_count	1.03e-01	0.000
followers_count:favorite_count	2.84e-09	0.000

made by the athletes. I used the "bing" sentiment dictionary to do this. The "bing" dictionary contains a list of 6,786 words that are either listed as negative or positive. From this, I found the number of positive and negative words and calculated the number of positive words divided by the total number of words found to be either positive or negative for each subset of data (political and non-political tweets). These two values were compared to analyze the overall sentiment of each respective dataset.

Lastly, I wanted to see if the models described above (the linear model and the regression model) would have significant predictive power. To test the effectiveness of the linear model, I looked at its R-squared value. For the logistic model, I first split up the dataset of all tweets into a train and test set, using 70% of the political tweets and the same number of non-political tweets to train it. I split the data in this manner so that the model would be able to better differentiate between political and non-political tweets, since approximately only one tenth of the dataset was deemed to be political and a normal random sample may not include enough political tweets to train the models well. To evaluate the logistic model, I produced a confusion matrix of its predictions.

Analysis

Shown below is the output of the linear model.

Factors which are statistically significant have a p-value less than 0.05, and these variables are minutes played, whether a tweet was political, the number of followers an account has, the number of favorites a tweet receives, and the interaction variable between followers and favorites. Most importantly for this study, the results demonstrate that whether a tweet is political is the most statistically significant variable in predicting retweets. This variable has by far the lowest p-value, and thus is most likely to have an impact on the retweet number. Thus, the linear model provides evidence that agrees with my first hypothesis since the model estimates that if a tweet is political it will have around 7,206 more retweets, on average. The output of the model exhibits that minutes played is the most important basketball statistic in determining the number of retweets a tweet receives. Surprisingly, however, it has a negative coefficient which leads to the conclusion that the more minutes an athlete plays the less retweets they will receive. I expected that players who receive more playing time would be more likely to gain more national exposure and thus twitter followers, which leads to more interaction with their tweets, but the model rejects this notion. After inspecting the dataset further, the reason for this unanticipated relationship is evident. If a player retweets another tweet, all retweets the original tweet receives will show up as retweets that the player's post receives. For example, if LeBron James retweets another user's post which has 500,000 retweets, the dataset marks this as James's tweet receiving those 500,00 retweets. Retweets generate a much greater number of further retweets compared to regular tweets or quote tweets, on average (7,717 to 341), and the average number of minutes played by users of retweets is significantly lower than that of normal tweets and quote tweets (974 to 1,129). These discoveries demonstrate why the basketball statistics part of the model are misleading due to how Twitter organizes the metadata of a retweet.

Shown below is the output of the logistic model.

term	estimate	p.value
(Intercept)	1.02e-01	0.000
PTS	2.89e-05	0.098
AST	3.65e-05	0.384
TRB	-5.74e-05	0.040
MP	7.68e-06	0.476
retweet_count	1.48e-06	0.000
favorite_count	4.58e-08	0.880
retweet_count:favorite_count	-2.00e-12	0.233

negative	positive	sentiment
672	1264	592

Just as in the linear model, factors which are statistically significant have a p-value less than 0.05, and these statistics consist of only total rebounds and the number of retweets. The results show that the retweet count p-value was around 1.88×10^{-30} which signifies that it is much more likely to have a real effect on the political nature of a tweet. Thus, having a higher number of retweets increases the probability that a tweet is political, on average. I want to note that although the p-value associated with the total rebounds variable is below 0.05 (around 0.04), it still is not extremely low. The data suggests that having a higher rebounding total reduces the chance a tweet from that player is political, which indicates that taller players are less likely to speak out about social justice issues. This finding is surprising to me, since big men have been just as publicly outspoken as all other players, and combined with the low but not super low p-value, it leads me to believe that it is more likely due to chance instead of actually being a significant result. These outcomes mostly contradict what I expected to happen. I thought more prominent players, ones who play more minutes and have better stats, would be more outspoken on social issues through twitter but the data suggests this is not the case. Furthermore, I expected the count of both favorites and retweets along with their interaction variable to have a significant impact on whether a tweet is political, however the data suggests that only the number of retweets does.

For the sentiment analysis, I found that 75% of the words that were in both the "bing" dictionary and in the text of the non-political tweets were considered positive, and 64% of the words in common between the "bing" dictionary and the content in political tweets was considered positive, as shown below.

Sentiment of Political Tweets

Sentiment of Non-political Tweets

These findings suggest my hypothesis was correct, however it is interesting to see how even political tweets are generally positive. A large portion of the criticism against athletes is that their calls for change are always negative complaints against the country, yet the data demonstrated that this is not the case at least in terms of their Twitter activity. One reason for this may be that in the past month almost all of the emphasis in the social justice sphere has been put on encouraging people to vote, and this type of encouragement usually has a positive connotation. It is possible that if I had gathered the twitter data a couple of months ago while the focus was more on the black lives matter movement that the sentiment might be more negative.

To test the predictive power of the linear regression, I examined the R-squared value of the linear model. The value is around 0.025, which means that about only 2.5% of the variability in the number of retweets a tweet receives can be explained by the model. This percentage is extremely low, so the model is not at all able to accurately predict retweets based on the factor variables listed above. This finding provides evidence against my hypothesis since I expected the model to be a good predictor of retweets. The lack of accuracy in the model is most likely due to the high variance in retweet count, as just over 10% of tweets (1,808) had zero retweets while the greatest number of retweets seen was 652,186. So although the model was able to

negative	positive	sentiment
2099	6065	3966

detect which factors have the greatest influence on the number of retweets, this did not lead to it being able to predict well.

Shown below is the confusion matrix consisting of the predictions made by the model on the test set, where "political" is whether a tweet is classified as political and "prediction" is what the model predicted a tweet should be grouped as, with zero being non-political and one being political.

```
##          prediction
## political    0    1
##         0 9630  465
##         1  401   66
```

The output demonstrates that the logistic model did not perform well in predicting whether a tweet was political. The main focus was on how accurate the model would be on identifying political tweets, and it only correctly classified 66 out of 467 tweets from the test set which were social justice related. So very similarly to the linear model, the logistic model could identify the factors that influence the political nature of a tweet the most but was unable to be a strong predictor of political tweets.

Conclusion

The most significant finding from this study exhibits that Twitter users interact more with political tweets compared to non-political tweets from NBA athletes. This is important because it solidifies that athletes do make an impact with the messages they put out through social media and indicates that the public does care about what athletes have to say on social justice and political issues. One of the main criticisms against athlete activism has been that few people are actually interested in their political views; however, this study proves this is clearly not the case. To illustrate this further, I wrangled the last 50 tweets posted by Laura Ingraham. The average number of retweets in this set of data is around 5,710, while the average number of retweets from the players' political tweets is 9,074. This shows that the average NBA athlete receives more retweets when their tweet is political than Ingraham's tweets do, on average, and gives a better idea of how far the activism of athletes through social media can go. The people want to hear what athletes have to say.

There are limitations to my study that may have affected the results. In the sentiment analysis, the sentiment of each set of tweets (political and non-political) was calculated through the total number of positive words compared to the total, rather than classifying each individual tweet as positive or negative and comparing the number of positive and negative tweets between the subsets of data. If I were to continue my analysis, I would incorporate this approach for a more accurate comparison. If I were to redo my project it would have been interesting to gather my data within the six weeks following the murder of George Floyd to capture more tweets that were about the black lives matter movement (since this was the original inspiration for my project), as by the time I began to wrangle data from Twitter it had already been around a month from when the NBA season ended and the social justice focus was much more firmly on the election.

References

Sources of Data

NBA 2019-2020 Player Stats: https://www.basketball-reference.com/leagues/NBA_2020_totals.html NBA Player Twitter Usernames: https://www.basketball-reference.com/friv/twitter.html

Additiontional Resources

Bakliwal, A. 2013. "Sentiment Analysis of Political Tweets: Towards an Accurate Classifier" http://doras.dcu.ie/19962/1/foster2013.pdf

Li, Z. 2014. "The Monetary Value of Twitter Followers: Evidences from NBA Players" https://aisel.aisnet.org/icis2014/proceedings/EconomicsandValue/20/

Molyneux, L. 2017. "Political Journalists' Normalization of Twitter" https://www.tandfonline.com/doi/full/10.1080/1461670X.2017.1370978