NBA Offseason 2018

Samad Patel

Abstract

The NBA Offseason is an incredibly exciting time. Trades, drafts, (the previous year's awards for some reason) and free-agents dominate the NBA media landscape and captivate the attention of fans around the world.

With each decision and change, fan reactions can be incredibly diverse. The purpose of this project is to analyze fan reactions to various free-agent decisions. The fan reactions will be taken from threads in the NBA subreddit (reddit.com/r/nba).

I will use sentiment analysis with the Natural Language Toolkit (nltk) to determine fan reactions. I will only use first-level comments, or, in other words, the comments at the top of the threads. The reasoning behind this is that top-level comments are more likely to be purely reactionary, whereas comments later in threads tend to diverge in topic.

I will run the comments through the sentiment analysis model to determine whether or not the reaction is positive or negative. Then I will weight the reaction based on the score of the comment (score = number of upvotes - number of downvotes). Thus, comments with a higher score are weighted more heavily since the score tends to indicate many other users agreed with the opinion.

The html/pdf version of this module will only demonstrate the output. Feel free to look at the jupyter notebook to view the code.

Part One: Creating the Model

To create the model, we need to feed it pre-classified words based on their positive or negative connotation. NLTK has a corpus called "opinion-lexicon" of positively and negatively connoted words. These are the words we'll use to train our model. Here are some examples from "opinion-lexicon":

```
Positive words: ['cheerful' 'meritorious' 'suitable' 'solidarity' 's
atisfactory']
Negative words: ['anguish' 'carelessness' 'backward' 'destitution' '
eyesore']
```

We'll now randomly split this data into a training and test set - 70% of the data for training, 30% to test with. Then we'll see what proportion of the test data the model correctly predicts. We're using a Naive Bayes Classifier.

```
The model correclty predicted 70.45% of the data
```

Not terribly bad for an untuned model. I'm not especially concerned for sky-high levels of accuracy, so we'll leave the model as is and continue on with the project.

Part Two: Collecting Data

We need to query reddit's API to select our data for each news thread. We'll be focusing on the following threads:

1. [Withers] LeBron signing with Lakers. (https://tinyurl.com/yaelyn95 (https://tinyurl.com/yaelyn95))
2. [Charania] Free agent DeMarcus Cousins has agreed to a deal with the Golden State Warriors. (https://tinyurl.com/y7kj8uyu (https://tinyurl.com/y7kj8uyu))
3. [Wojnarowski] Paul George has committed to sign a deal with the Oklahoma City Thunder, league sources tell ESPN. (https://tinyurl.com/yav7akjz (https://tinyurl.com/yav7akjz))

LeBron to the Lakers

Here are the top three comments on that thread:

```
b"as a lakers fan since july 2018 i can't put into words how much th
is means to me"


b'Comments moving so fast no one will know I love my wife '


b'Does this hurt his chances of going to Houston?'
```

Boogie to the Warriors

Here are the top three comments on that thread:

```
b'_R.I.P Lakers bandwagon_\n\n_7/1/2018 to 7/2/2018_'


b'Kill me\n\nEdit: 5.3mil..... I fucking hate everything '


b'CURRY/KLAY/KD/DRAY/DMC\n\nSHUT THE WHOLE DAMN LEAGUE DOWN'
```

Who'd have thought the internet would be angry at the Warriors for adding another All-Star??? But does the model reflect this anger? We'll find out shortly.

PG13 Staying in OKC

Here are the top three comments in that thread:

```
b'Delete thi- Wait no this one is real'
```

```
b'This is why you trade for a Superstar even if his contract is expi
ring, good for OKC.'
```

```
b'I know this is a thread for one-line reactions and all that but...
I am really, really surprised. I thought he was *so* gone after that
press conference and that awful first round loss to the Jazz. This s
hould really put paid to a lot of the Westbrook hot takes that are s
o common today, and it should also show the value of trading for a p
layer and taking a gamble, because it might turn out a lot better th
an you think.\n\nCongrats, OKC! Another year to make it work. \n\nED
IT - another four years, damn'
```

Part Three: Implementing the Model

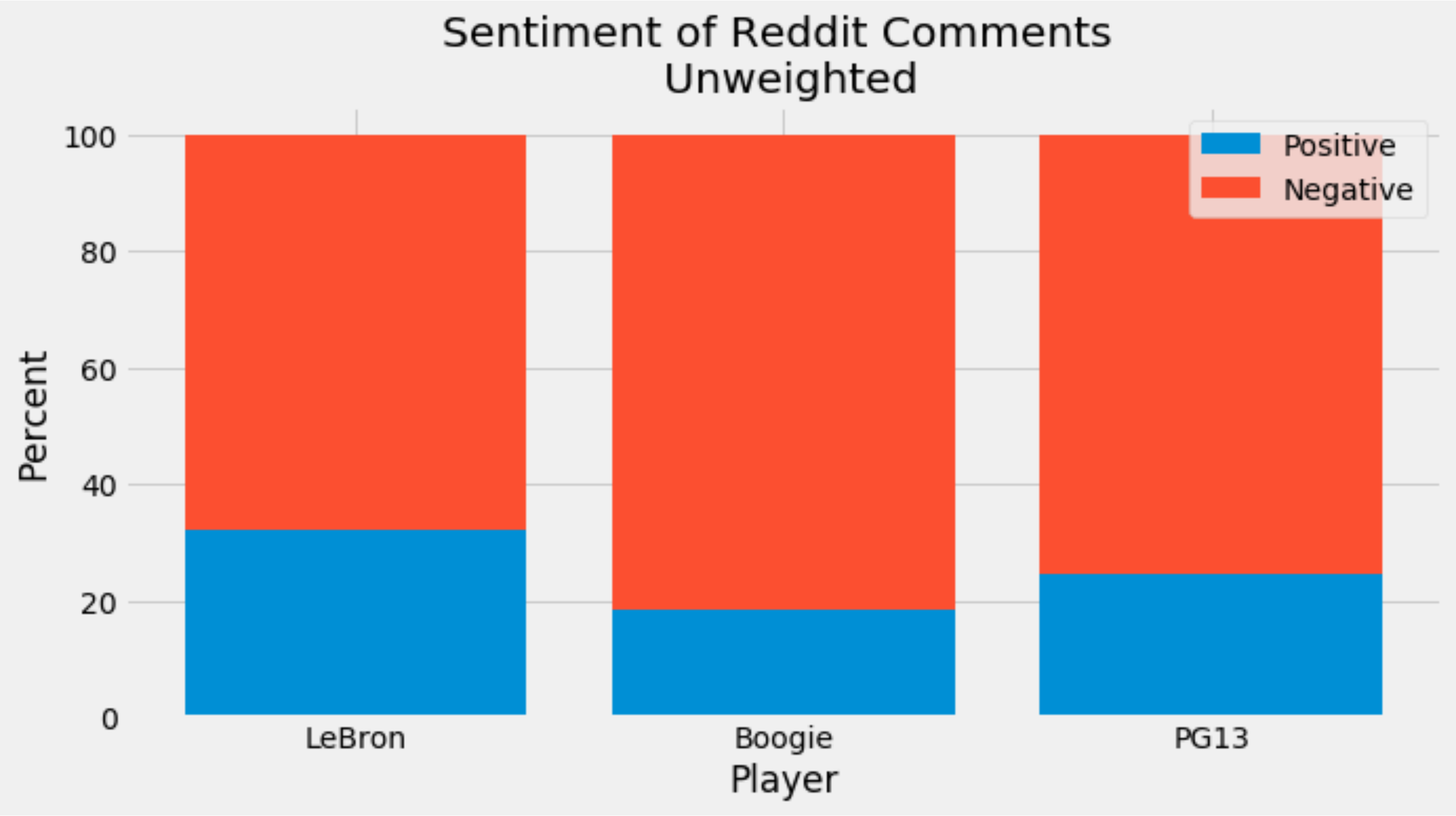
This section formats the data in the way the model understands and then spits out the class predictions. We'll break it down into two parts - one model assuming that each comment should be weighed equally, and another that assumes higher-scored comments have greater weights.

Unweighted Model

```
About 67.86% of comments regarding LeBron's move were negative.

About 81.42% of comments regarding Boogie's move were negative.

About 75.23% of comments regarding PG's move were negative.
```



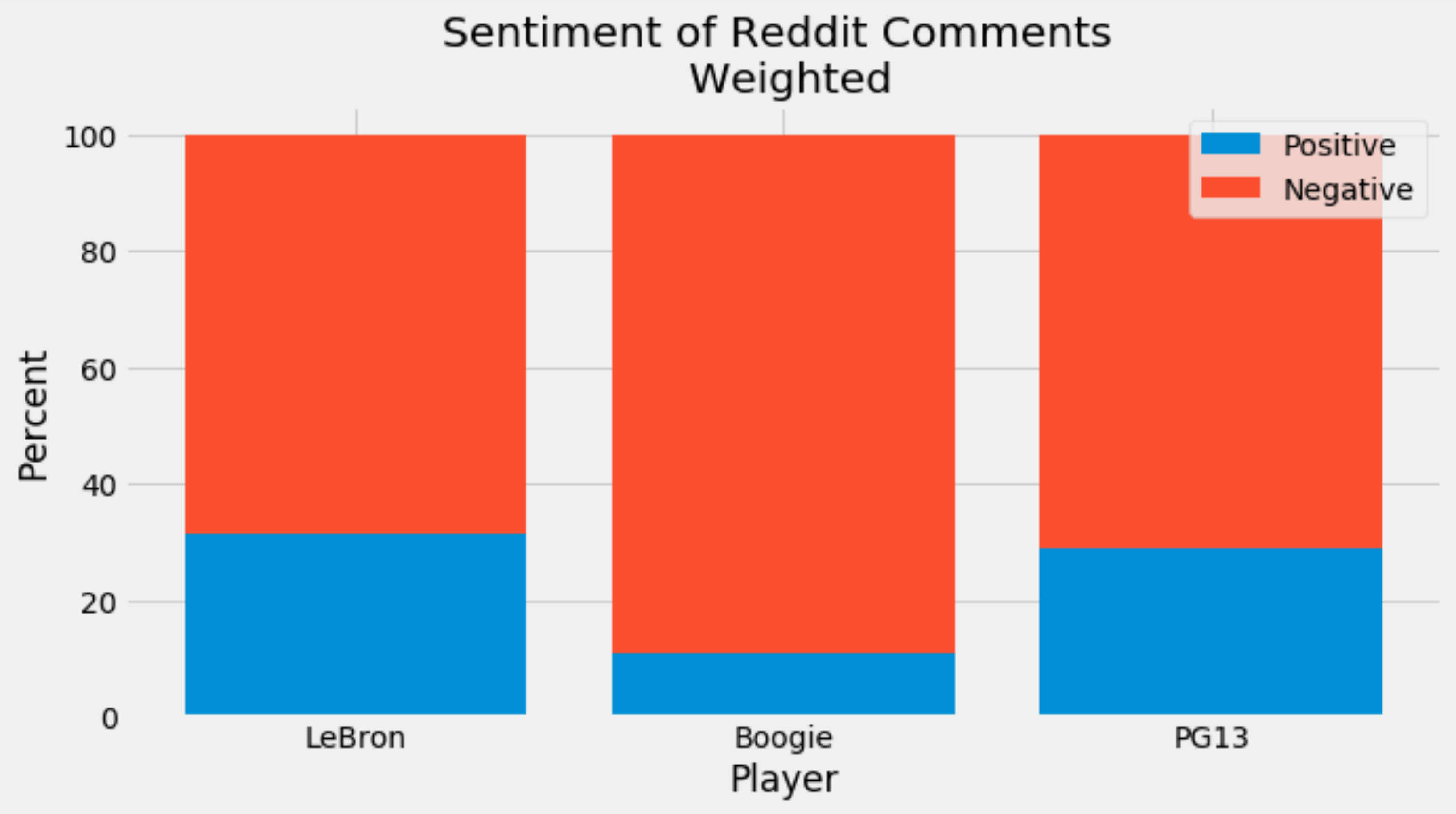
Weighted Model

First, we need to use the API to determine the score for each comment. Then we'll determine the weight as the proportion of each score over the sum of all scores.

```
About 68.61% of comments regarding LeBron's move were negative.


About 89.24% of comments regarding Boogie's move were negative.


About 71.13% of comments regarding PG's move were negative.
```



Part Four: Improved Models

From skimming the comment threads, they don't seem to actually be this negative. It turns out that the model classifies any words it's unfamiliar with as negative, which leads to an emormous overestimation of negative words. The words the model is trained on are all properly spelled and punctuated, which is not representative of the nature of reddit coments.

There is no available corpus in the NLTK library for reddit comments, so let's try using the twitter_samples and sentence_polarity.

Twitter Samples

```
The accuracy of the twitter_samples model is 75.51515151515152.
```

```
About 60.71% of comments regarding LeBron's move were negative based
on the twitter_samples model.
About 66.06% of comments regarding PG's move were negative based on
the twitter_sampels model.
About 69.03% of comments regarding Boogie's move were negative based
on the twitter_samples model.
```

Sentence Polarity

```
The accuracy of the sentence_polarity model is 75.48295454545455.
```

```
About 60.71% of comments regarding LeBron's move were negative based
on the sentence_polarity model.
About 66.06% of comments regarding PG's move were negative based on
the sentence_polarity model.
About 69.03% of comments regarding Boogie's move were negative based
on the sentence_polarity model.
```

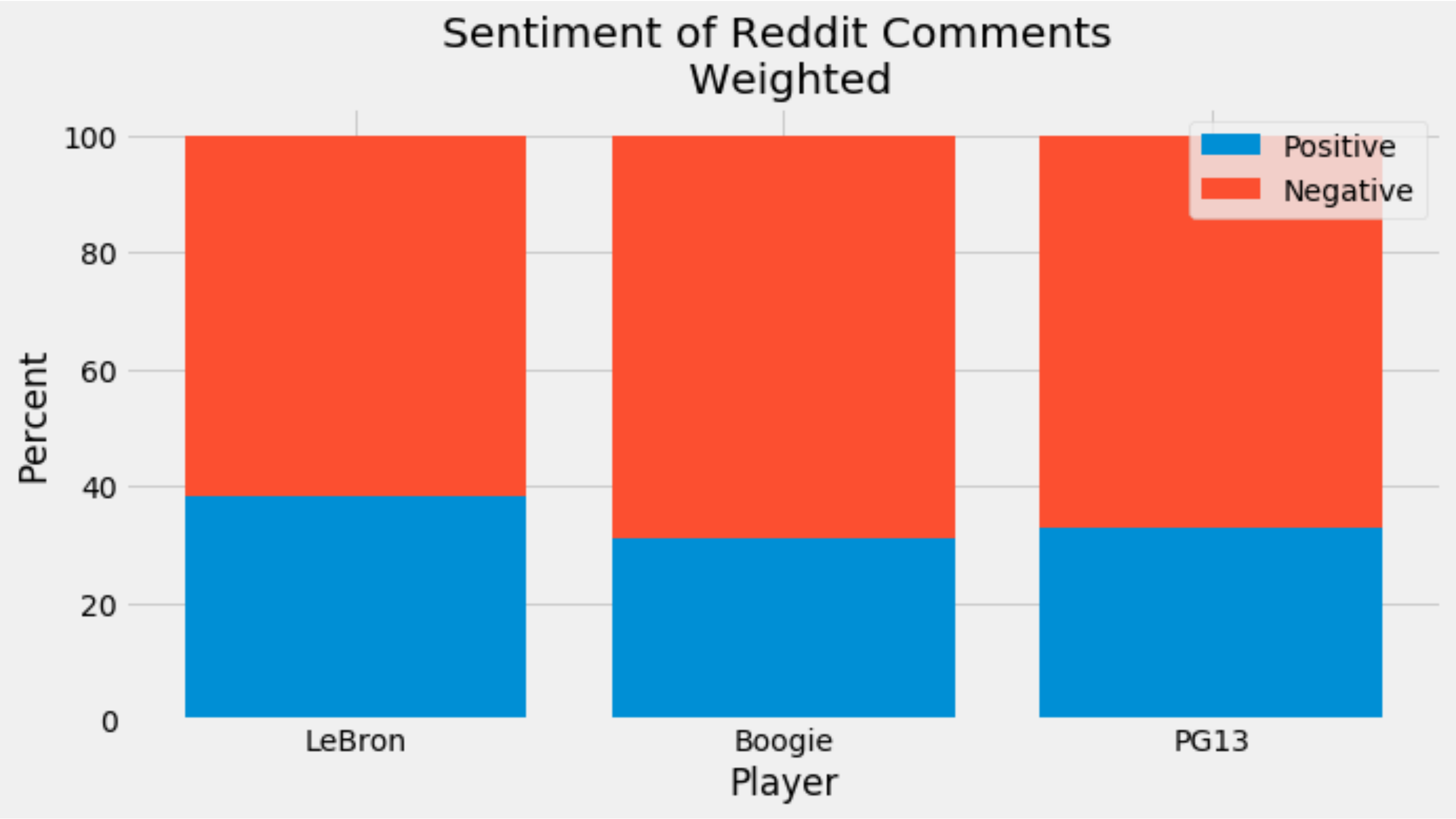
Conclusion:

It's clear that the twitter samples and sentence polarity corpuses give better results than the opinion_lexicon corpus. They give very similar results, so the choice here isn't especially substantial. I'll select the sentence_polarity model since the sentences here are in longer form than the twitter comments, and so they therefore mirror reddit comments better.

Part Five: Final Model/Final Results

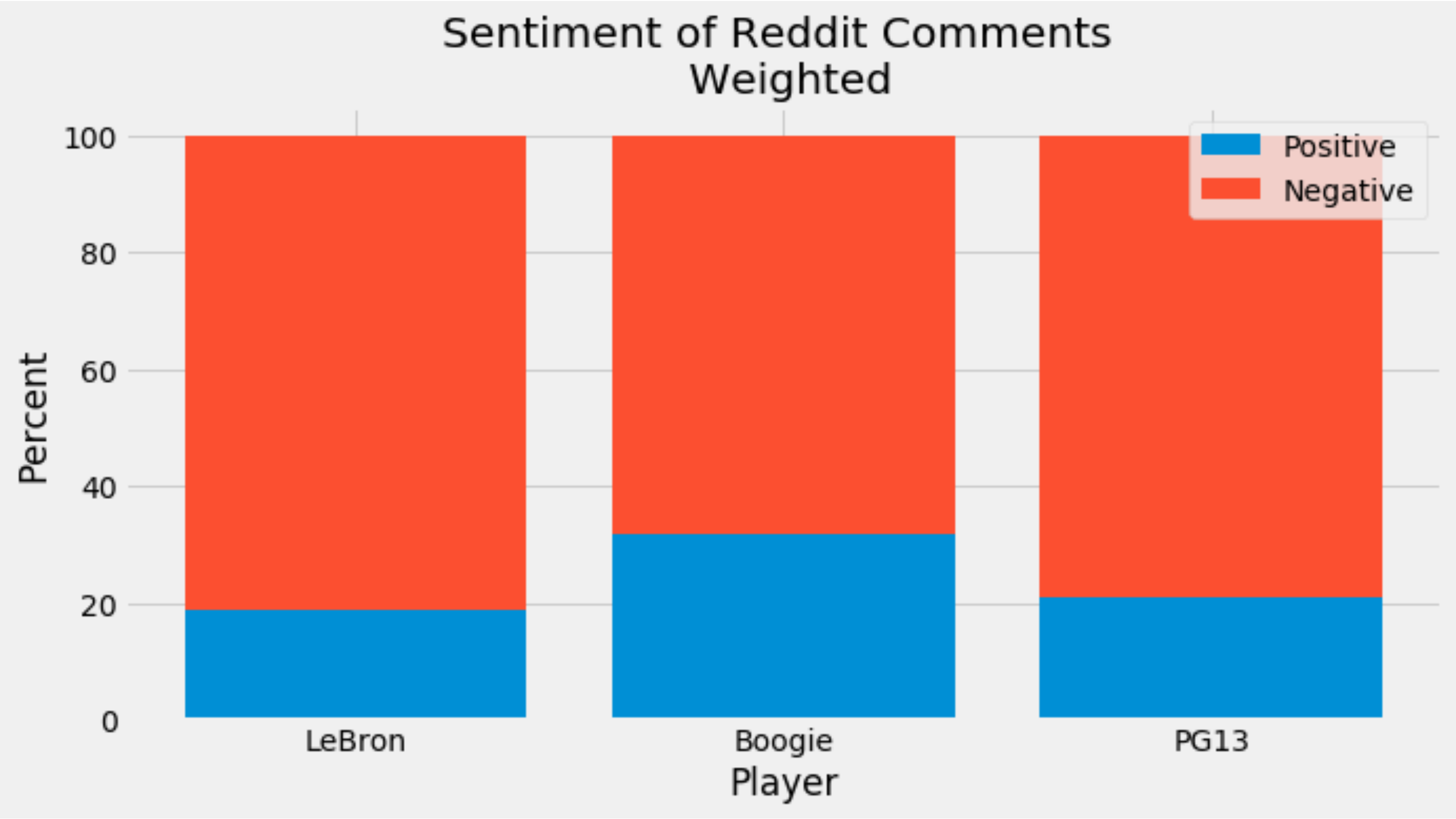
Unweighted Model

The results are shown in the previous section are reflected in the plot below.



Weighted

```
About 81.11% of comments regarding LeBron's move were negative.
About 79.02% of comments regarding PG's move were negative.
About 68.27% of comments regarding Boogie's move were negative.
```



Part Four: Conclusions

Interpretation of Results

In the unweighted model, the order of negativity from least to greatest is LeBron, PG13, and Boogie. In the weighted model, the order reverses - Boogie, PG12, LeBron. How could this be?

Context is incredibly important here as far as interpreting what negativity truly means. It doesn't inherently mean that users are lambasting the decision itself, or the player themselves. For the Paul George decision, much of the thread focused on how bad Lakers fans feel or how unexpected the decision was given the expectation that PG would go to the Lakers. For LeBron, many users were incredibly shocked, using words like fuck and shit frequently. The only thread analyzed where the majority of negative comments were in fact lambasting the move itself was with DeMarcus Cousins, which was to be expected given the narrative of a fifth All-Star joining one of the best teams of all time. But still, many of these comments were expressing major shock, again using words like fuck (outside of "fuck the warriors/fuck boogie").

Because there are also high rates of misclassification in the model, misclassified entries can dramatically skew the weighted model. Therefore, I find it best to select the unweighted model.

Limitations

There are also big shortcomings in the model. First, the data wasn't trained on reddit comments, so the model and accuracy are hardly representative of the comments we're passing through. Thus, we have a much higher proportion of negative words, becuase the classifier determines any word it doesn't understand is negative.

Second, it can't determine context from certain neutral phrases. For example, "Holy fuck!" or "Holy shit!" show up very frequently, but both "fuck" and "shit" are counted as "negative" words. So "Holy fuck this is amazing!" can be considered negative if fuck is weighted more negative than amazing is weighted positive.

All in all, the model needs to be trained much more to derive stronger conclusions from these sorts of threads. As is, however, they give us good insight on an ordinal level. We can certainly conclude that LeBron's decision was the most well-received move compared to the others. The extent can be determined with future analysis.