Philadelphia loves to hate Harper

IST736 Text Mining

Martin Alonso

2019-03-19

Introduction

For many seasons, Twitter has served as a thermometer for sport fans to either praise or lambast many professional athletes. For baseball, this is no different. Over the course of the 2018-2019 Major League Baseball (MLB) offseason, fans have shown displeasure at how slow player transactions have occurred. One especially stands out: the signing of, or lack off, Bryce Harper.

At 26 years old, Harper is one of the youngest player to have ever reached free agency in MLB. Many regarded Harper, a prodigy of the sport, to command record offers for his services. Contracts paying up to $400 million dollars over ten years where not discounted. But, once November 1st, 2018 came around, the start of the baseball offseason, news about Harper was slow. Aside from a 10-year $300 million dollar offer by the Washington Nationals, which was rejected, only news of teams meeting with Harper reached to fans.

Fans of the Yankees, Dodgers, Cubs, and Phillies grew restless, expressing their displeasure about both their teams' inability to sign Harper, and Harper's alleged egoism. Twitter either praised him in want for news, egged him to sign quickly, or spewed hateful messages about the glacier-like speed that negotiations took place.

Finally, on March 1st, 2019, Harper signed a record 13-year $330 million dollar contract with the Philadelphia Phillies and Twitter erupted. Messages from all fans displaying either joy or hate invaded Twitter.

***

The objective of this project is to analyze twitter sentiment towards Bryce Harper over the course of MLB's offseason. Given how long it took to sign him, and how many teams fought for his services, we want to find out how people from the cities vying for his services feel about him.

The hypothesize is that, as time passes, and one team comes out as a favorite candidate to sign Harper, fans of said team, the Philadelphia Phillies, will have a more positive sentiment towards Harper while fans of other teams will be negative towards him. However, we must be mindful of the sports adage that no fans are more critical than Philadelphians, they will hate an athlete even if he has yet to play a single professional game for a Philadelphia franchise.

Methodology

Between November 1st, 2018 and March 1st, 2019, 327,148 tweets mentioning Bryce Harper were sent out into the Twittersphere. These tweets were downloaded using the Twitter developer API to obtain tweet ids, usernames, timestamp, and tweet content. However, crucial to this project was getting location from where the tweet was sent. Since this

information was not returned by the API, Python's twitterscraper package was used to parse through these tweets and obtain location data.

Given that the twitter location data is able to be manipulated by the user, these data points needed to be handchecked to obtain as accurate a picture as possible about the real location these tweets were sent out. Once this was managed, we could group the tweets by location and proceed with the sentiment analysis, using the Python package NLTK and Vader sentiment analysis.

After these data were cleaned and analyzed, the texts were both stemmed and lemmatized. We then built four models using TF-IDF and Count Vectorizers to parse the content of each tweet; finally passing the processed data through a Multinomial Naive Bayes and Linear Support Vector Machine algorithms.

1. Data Munging

   To obtain the data, the Python packages tweepy and twitterscraper were used. The tweepy package allows a user to connect to the Twitter developer API, capable of scraping tweets about any subject. Unfortunately, the tweepy package and API only allow a user to scrape data up to one week old; an alternative to obtaining the data was thus found.

   The twitterscraper package allows users to select a subject, Twitter user, or dates to look for tweets on any subject. For this project, we focused on tweets mentioning Bryce Harper during the MLB offseason. Though the data was obtained fairly quickly and successfully parsed, location data was missing.

   Fortunately, tweepy allows users to input a tweet's id, returning the user, tweet, and, more importantly, location. Using IBM Watson Studio and PySpark, the 327 thousand tweets were parsed, obtaining location data for 68 percent of the tweets. The remaining 32 percent returned null ids.

   The data needed further cleaning, as location data was not normalized. Though there were obvious mentions of cities, states, and countries from around the world, there were also numerous references to places non-existent or childish references. The data was tried to be cleaned as granularly as possible but, ultimately, around 45 percent of the data was successfully cleaned. That still left us with a database of 117,254 tweets to work with.

2. Data Analytics

   The first thing we wanted to do was check how many tweets we had from each city. From the cleaned tweets, we had 804 cities from around the world, though the majority were located in the US (table 1).

   At first glance, nearly half of the tweets are coming from Philadelphia, which makes sense given the Harper ultimately signed with the local team. But this also weighs heavily against other teams and cities. Using geographic knowledge, cities and districts were grouped together (table 2) to try and gain more parity among cities.

Though this was not feasible, we did get the data to be more representative of other cities not Philadelphia (table 3).
Of the 15 cities shown, the top ten contained 86 percent of the tweeted data, so we decided to move on with these cities: Philadelphia, New York, Chicago, Washington DC, Los Angeles, San Francisco, San Diego, St. Louis, Las Vegas, and Boston.

Table 1: Top 15 cities, pre-cleaning

Rank	City	Tweets
1	Philadelphia	56,762
2	New York	8.516
3	Chicago	8,330
4	Los Angeles	5,991
5	Washington D.C.	5,551
6	San Diego	3,260
7	San Francisco	2,583
8	St. Louis	2,522
9	Las Vegas	1.695
10	Boston	1,231
11	Atlanta	1,125
12	Brooklyn	1,107
13	Houston	1,091
14	Phoenix	868
15	Seattle	670

Table 2: City groupings

City	Contains
New York	{New York, Bronx, Brooklyn, Staten Island, Long Island, Queens}
San Francisco	{San Francisco, San Jose, Sacramento, Oakland}
Los Angeles	{Los Angeles, Santa Monica, Hollywood, Long Beach, Huntington Beach}
Washington DC	{Washington DC, Richmond, Baltimore, Alexandria}

Table 3: Top 15 cities, post-cleaning

Rank	City	Tweets
1	Philadelphia	56,762
2	New York	10,659
3	Chicago	8,330
4	Washington D.C.	6,662
5	Los Angeles.	6,114
6	San Francisco	3,696
7	San Diego	3,260
8	St. Louis	2,522
9	Las Vegas	1.695
10	Boston	1,231
11	Atlanta	1,125
12	Houston	1,091
13	Phoenix	868
14	Seattle	670
15	Miami	615

Having grouped the data to a more manageable count, the second question we wanted to answer was how many tweets were mentioning Harper were being tweeted each day (figure 1). On average, 840 tweets were being sent out daily. But, this information is skewed, considering that more than 40 thousand tweets were sent out the day Harper signed with the Phillies.

The data were split into before February 27, 2019 and after the date, to compare how each city was behaving (figures 2 and 3) and some very interesting patterns surfaced.

Evidently, as teams were rumored to be in the running, tweets from that city increased. A brief rumor regarding the Chicago Cubs spikes tweets from Chicago around mid-December. After that, Los Angeles becomes a front runner but quickly dies out.

It's not until early January that Philadelphia starts appearing as a viable candidate and maintains heavy tweets usage from here until the actual day that Harper signs, when

over 40 thousand tweets are sent out about him. Clearly, we can see that Harper being linked to a team drives tweets from said city's team.

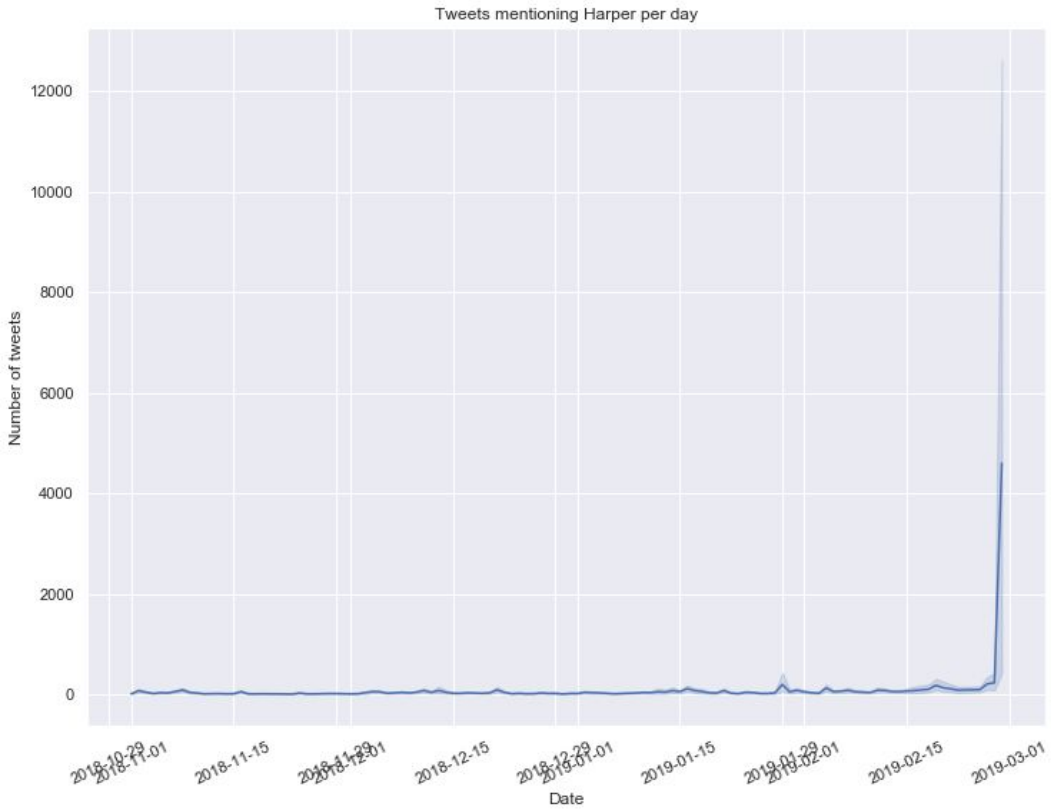


Figure 1: Number of Tweets mentioning Harper per day

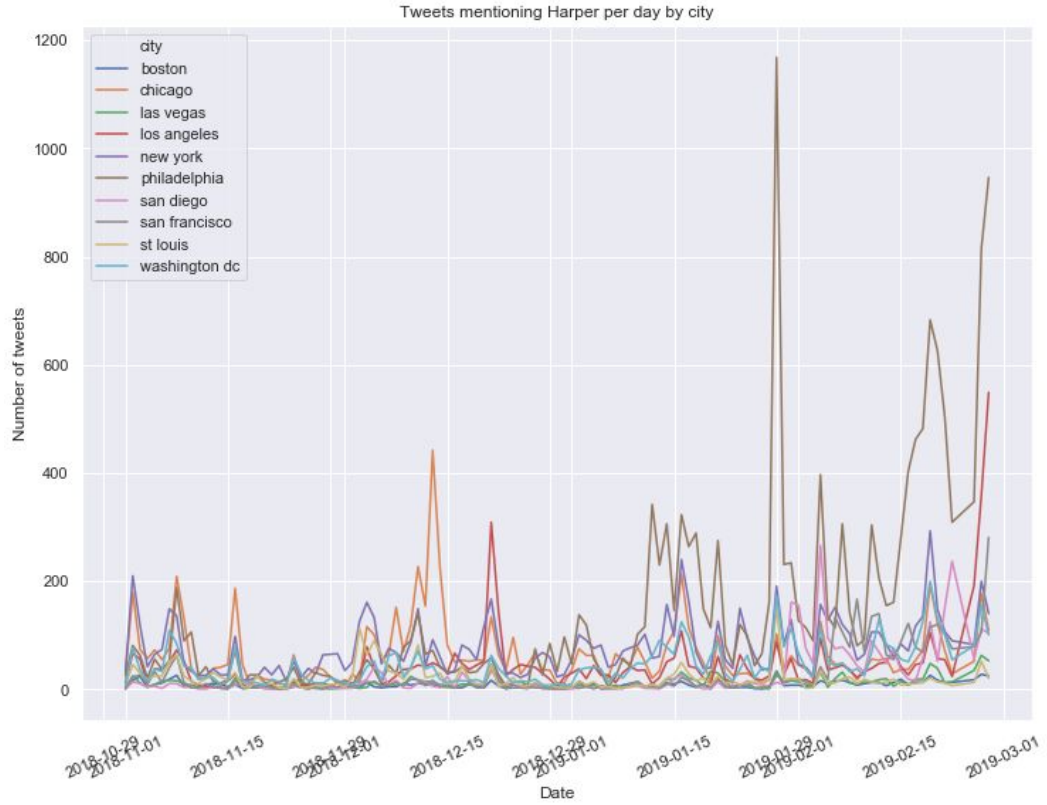


Figure 2: Tweets between 2018-11-01 and 2019-02-27 mentioning Harper by city

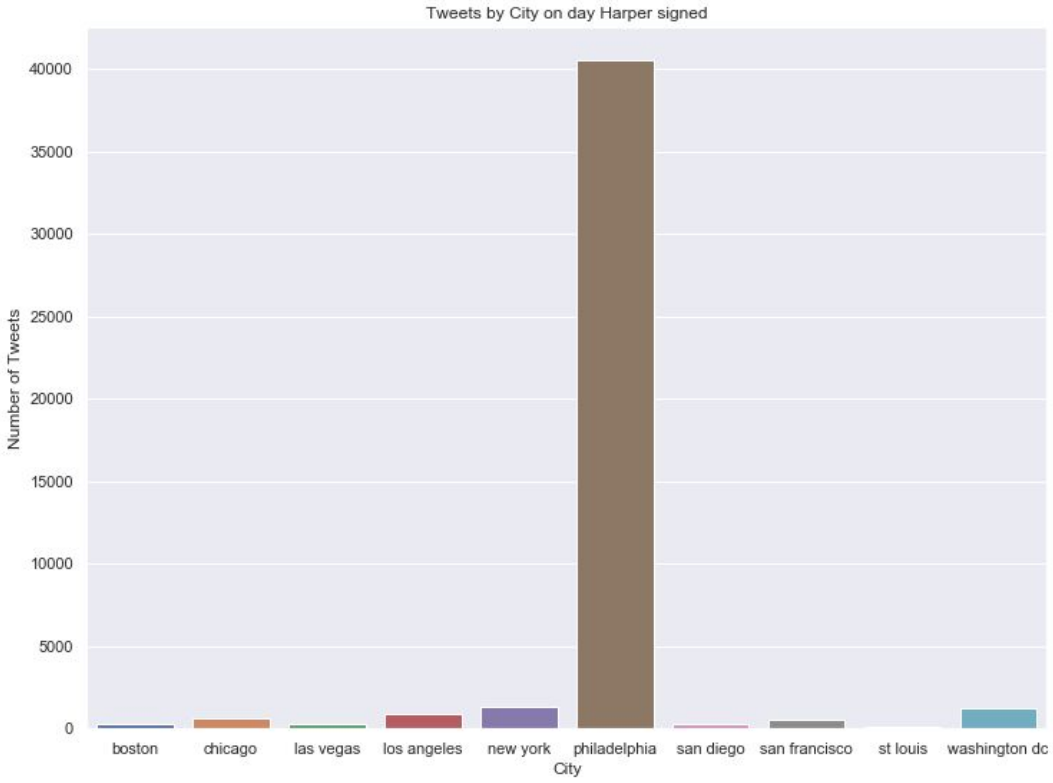


Figure 3: Tweets by city mentioning Harper on 2019-02-28

With this relationship between timeline and city established, it was time to analyze the actual tweets.

From the data, tweet id, city, text, and location were kept. The text was of particular interest in this case. The information was cleaned, removing emojis, special characters, and stop words; then lower-casing every character. Numbers were not removed as these could potentially refer to baseball stats used to support or detract Harper's case as a great and valuable player.

From the NLTK package, the Porter Stemmer Word Net Lemmatizer functions were imported, as the data needed extra cleaning before doing any analysis. The texts were vectorized, passed through the functions, and reconstructed as partial sentences containing the text data (table 4).

The next part required loading the Vader Sentiment analyzer, again from the NLTK package, and identifying the sentiment of the text. Both the stemmed and lemmatized versions of the texts were passed through the sentiment classifier, with those texts with a compound score greater than 0.2 being classified as positive and those with a score below -0.2 being classified as negative (table 5). The sentiment analyzer did a pretty good job. From ten tweets chosen at random, nine presented a sentiment classification in accordance with the score given by the Vader analyzer.

Table 4: first five observations of text, stemmed text, and lemmatized text

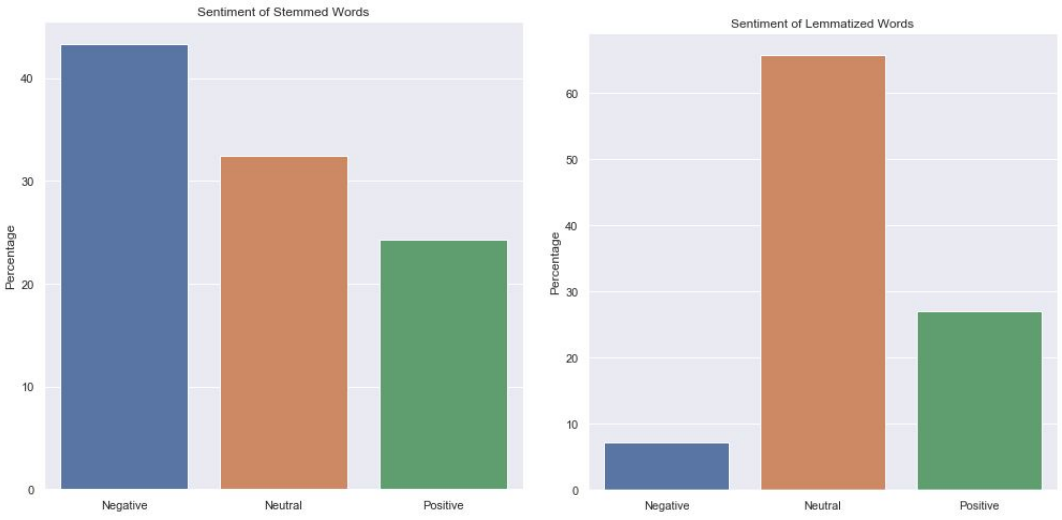
Observation	Text	Stemmed	Lemmatized
1	bryce harper with the cubs	bryce harper cub	bryce harper cub
2	white sox bryce harper	white sox bryce harper	white sox bryce harper
3	what s bharper3407 deing make up your mind bruh	bharper3407 make mind bruh	bharper3407 make mind bruh
4	best team with the biggest wallet will sign th..	best team biggest wallet sign best free agent	best team biggest wallet sign best free agent ....
5	haha you d fit right in	haha fit right	haha fit right

Table 5: Vader Sentiment

Observation	Text	Sentiment Score
1	Bryce harper cub	Neutral
2	White sox bryce harper	Neutral

3	bharper 3407 make mind bruh	Neutral
4	Best team with the biggest wallet sign best free agent...	Positive
5	Haha fit right	Positive
6	Baseball history backs youth performance good...	Positive
7	Bryce harper reportedly told friends living ph...	Positive
8	Pic twitter com g7t9eevjh1	Neutral
9	Bryce harper villain yankee needed long time...	Negative
10	Bryce harper future yanke pic twitter com 69b...	Neutral

Furthermore, we decided to check the distributions of the stemmed and lemmatized tweets (figures 4 and 5). Clearly, there's a different story to tell using stemmed texts than lemmatized this. For the sake of this project, we'll move forward using stemmed text sentiment.



Figures 4 and 5: Tweet sentiment for Stemmed and Lemmatized texts

The question now becomes, how does each city feel about Harper overall. Given that Philadelphia signed Harper, we'd expect them to be more supportive of him. However, this is not true (figure 6) as the majority of tweets from Philadelphia are negative, while the rest of the cities are mostly positive towards Harper.

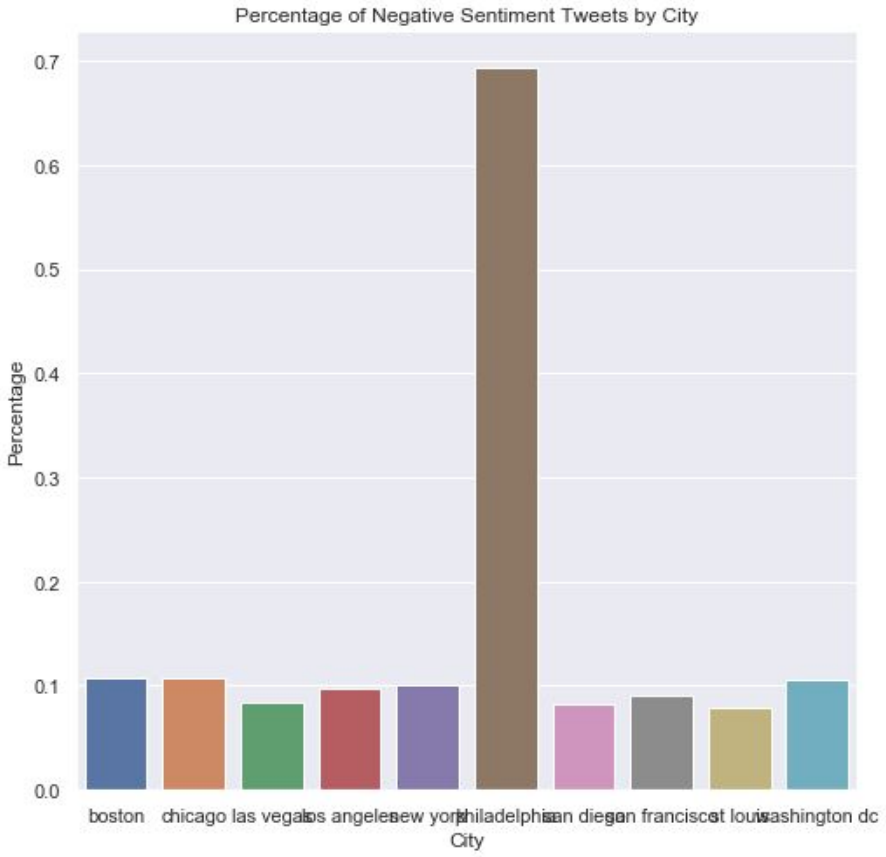


Figure 6: Negative sentiment toward Harper by city - all Tweets

There is a clear bias towards Harper that pertains only to the city of Brotherly Love. When did this happen? Our hypothesis was that Philadelphia would welcome Harper openly but that is clearly not the case. Judging by the word cloud (figure 7), Philadelphia fans were growing restless with Harper not signing and Philadelphia overspending to obtain his services.

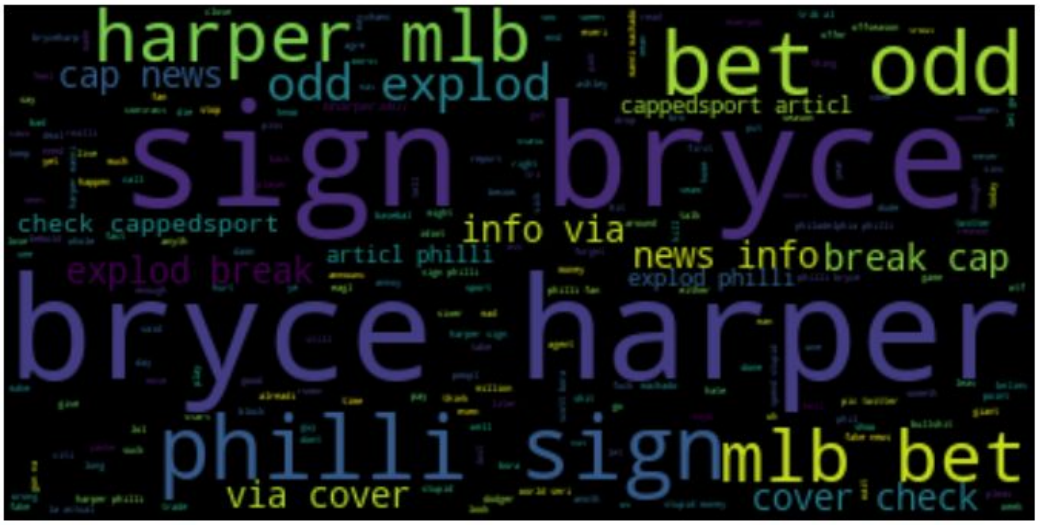


Figure 7: Word cloud of Negative sentiment tweets from Philadelphia

However, negative sentimentality never grows against Harper. Between the start of the offseason and the day prior to signing his contract, the average percentage of negative sentiment tweets hovers around 10 percent. Only when he signs his contract with Philadelphia does the negative sentiment towards him grow to almost 70 percent, and that's only in Philadelphia (figure 8).

With the data analyzed, we can start building a model. Our assumptions going into this next part is that any negative tweet will mostly be attributed to Philadelphia. However, because the Philadelphia tweets make more than 50 percent of this subset, the model may be heavily skewed towards this city.

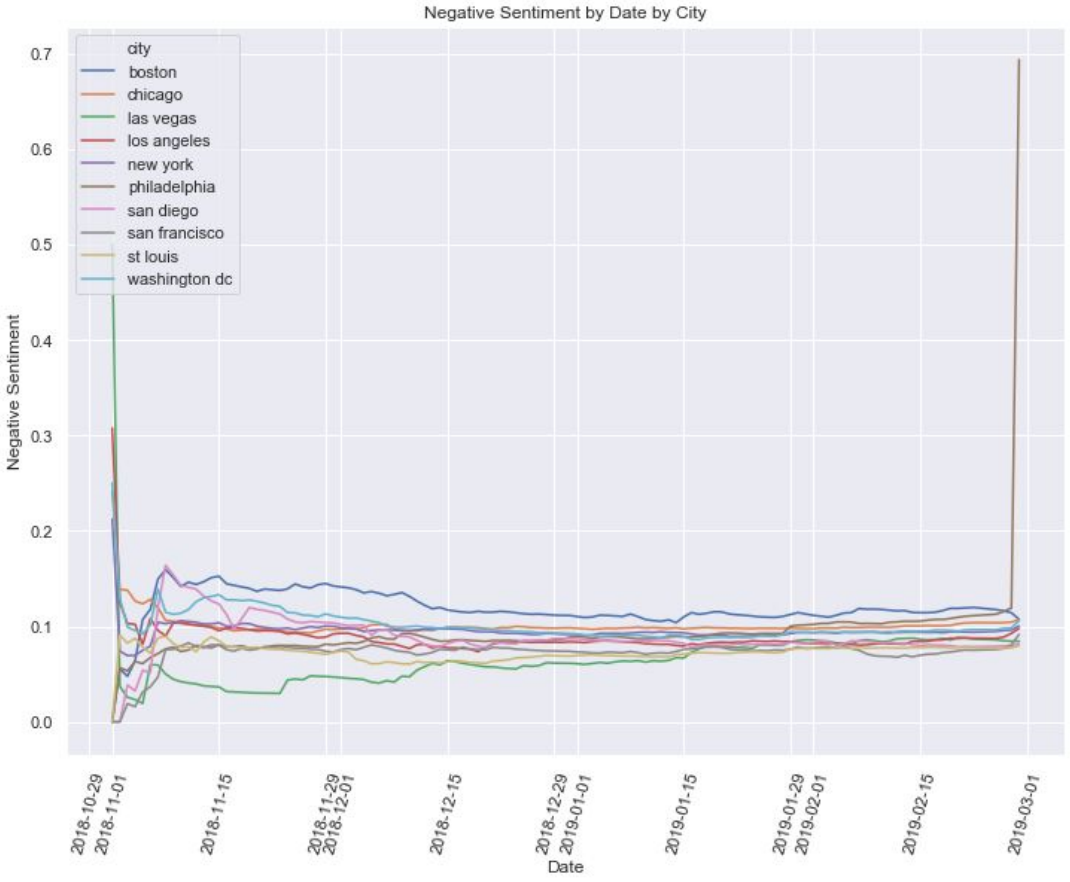


Figure 8: Percentage over time of Negative Sentiment tweets about Harper

3. Data Modeling

We continue using the stemmed texts for the model construction. The objective of the model is to, using the stemmed text from the tweets, correctly identify and predict from which city the tweet was sent.

For this data subset, we will build four models using Multinomial Naive Bayes and Linear SVM. For each algorithm, we'll use both TF-IDF and Count Vectorizers, using unigram and bigram word detection, minimum word frequency of five words, l2-norm regularization, and, for the Count Vectorizer, binary identification and C-penalization equal to 1.0.

Using these features, the models were trained using 80 percent of the dataset and measured using sklearn's accuracy score. The results of the Multinomial Naive Bayes were not encouraging, as the model barely managed to reach 60 percent model accuracy. The Linear SVM, on the other hand, managed to surpass 80 percent accuracy.

The Multinomial Naive Bayes was tuned to try and improve performance but, unfortunately, the gain was good but not great, surpassing the 65 percent mark (table 6), but still very far from the Linear SVM.

Table 6: Model performance on training set

Model	Accuracy Score (%)
Multinomial NB - Count Vectorizer	65.8
Multinomial NB - TF-IDF Vectorizer	67.5
Linear SVM - Count Vectorizer	82.4
Linear SVM - TF-IDF Vectorizer	86.7

Satisfied with these results, we'll move on to test the models on the remaining 20 percent of the data, the test set, and four new tweets that I wrote.

Results

The four models were ran on the test set, but the results varied compared to the training sets (table 7).

Table 7: Model performance on test set

Model	Accuracy Score (%)
Multinomial NB - Count Vectorizer	61.5
Multinomial NB - TF-IDF Vectorizer	62.8
Linear SVM - Count Vectorizer	70.6
Linear SVM - TF-IDF Vectorizer	70.6

Though the Multinomial Naive Bayes models had very low accuracy scores during training, the results on the test set were not that far off, meaning that the models were not suffering from over or under-fitting.

The same, nevertheless, can not be said of the Linear SVMs. Both these models outperformed the MNB algorithms during training, and though they again outperformed them on the test set, it is clear that the algorithm overfitted on the training data.

Focusing solely on the Linear SVM - TF-IDF model, we created a confusion matrix to identify were the model was missing and were it was most accurate (table 8). Though all over the place, the model correctly identifies 86.7 percent of tweets coming from Philadelphia. It also correctly identifies tweets coming from New York and Chicago. However, it fails to distinguish Boston and Las Vegas from any other city, and performs neutrally for San Diego, San Francisco, and St. Louis.

Table 8: Linear SVM - TF-IDF Confusion Matrix (horizontal: predicted; vertical: actual)

	Bos	Chi	LV	LA	NY	Phi	SD	SF	StL	Was
Bos	47	20	7	13	88	33	11	4	5	13
Chi	36	959	16	53	307	181	25	29	27	51
LV	15	24	89	41	81	62	9	16	1	11
LA	40	70	24	473	270	238	19	40	22	51
NY	81	95	17	65	1.348	322	29	41	21	88
Phi	111	142	44	157	614	9.847	77	82	48	142
SD	14	36	10	45	101	113	300	27	4	18
SF	24	32	9	41	116	102	19	356	5	27
StL	14	35	7	18	98	49	8	7	256	28
Was	44	61	17	53	278	248	31	32	28	584

Finally, I wrote four tweets that the model has not seen, and ran these through the four models to identify from which city these tweets came from (table 9), along with my intent on which cities I was thinking about when writing these tweets.

Table 9: Validation tweets

Tweet	Intent	MNB-CV	MNB-TFIDF	SVM-CV	SVM-TFIDF
Harper shouldn't sign with the Phillies! He's terrible	Philadelphia	Philadelphia	Philadelphia	Philadelphia	Philadelphia
"13 years $330 million?!?!?!. Steinbrenner's could've easily topped that	New York	New York	New York	Washington	New York

Bring Bryce to Hollywood!	Los Angeles	Los Angeles	New York	New York	New York
F@ck him! He doesn't deserve that money! Very inconsistent player	Philadelphia	New York	New York	New York	Washington

Despite our best efforts, it appears that, on this small validation set, the MNB-CV does the best job, successfully identifying three out of the four intents; the other three models barely getting two right.

Conclusions

After working through the data analysis section, we find that the initial hypothesis, Philadelphia fans were happy and excited to sign Harper, was incorrect. Despite a vast number of tweets mentioning Harper that originate from Philadelphia, it is the fans from other teams and cities that are more excited about him than actual people in Philadelphia.

Regarding the model, we can see that the results on the test set for the Linear SVM model are great, but can still be improved upon with further fine-tuning. However, it is of major concern that the model is overfitting on the training data. The same, cannot be said about the Multinomial Naive Bayes model. Though the model performs poorly compared to the SVM, the model does not suffer from the same overfitting problem. This means that we can still fine tune the vectorizers and algorithm to improve the accuracy of the model while, hopefully, not increasing the gap between the accuracy of the training and testing sets.

Finally, though the Linear SVM model using TF-IDF vectorizer performs best out of all the models, when it comes to the validation tweets, it turns out, again, that the Multinomial Naive Bayes is more accurate. Granted, the sample size of this validation set is only four tweets written by myself, so these results must be taken with a grain of salt. A better validation set would be to grab a larger sample of tweets from the areas tested and compare the results of the model.

This model and project still has many improvements to build upon; I have no doubt that it can lead to better understanding of how fan bases react and feel about players, while also helping us identify ideas behind each fanbase, capable of letting us identify their rooting interests. Further model calibration, plus obtaining data about other players is key to understanding and comparing how fans and teams feel about certain players, whether it's the best player in baseball or a lowly-franchise addition.

References

- https://www.thegoodphight.com/2019/2/28/18230001/a-complete-social-media-recounting-of-the-bryce-harper-saga-phillies-hot-stove

- https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f
- http://blog.chapagain.com.np/python-nltk-sentiment-analysis-on-movie-reviews-natural-language-processing-nlp/
- https://www.datacamp.com/community/tutorials/simplifying-sentiment-analysis-python
- https://www.pingshiuanchua.com/blog/post/simple-sentiment-analysis-python?utm_campaign=News&utm_medium=Community&utm_source=DataCamp.com
- https://developer.twitter.com/en/dashboard
- https://www.twitter.com/

Special thanks to the Baseball Prospectus Stats team for their valuable feedback for this project and helping point out some potential pitfalls and improvements to the model. Special thanks to Dan Brooks for suggesting the idea and Sean O'Rourke for his insights into the Philadelphia Baseball Twitter landscape.

The entire project can be found in the following GitHub page.