Don't expect the public to look at your stadium any different as you come close to finishing it

tl;dr

If you're in the business of building stadiums, public sentiment towards your project will most likely not change as you near completion. This project analyzed the content of over 274,000 tweets from the last three years that related to the ten largest stadiums under construction in America to explore the relationship between public sentiment of stadiums under completion and the months until completion.

SASHA NOVITSKY
Economics , Claremont McKenna College
AFZAL PATEL
Computer Science, Pomona College

Dec. 19 2019

Citation:
Novitsky & Patel, 2019

The Problem

Within the last few decades, new sports stadiums seem to be popping up every day. While many of these stadiums are financed privately by individuals and corporations, the majority are financed by a mix of public and private funds. But what do the citizens really think of these costly stadiums? [1]

Overview

A dataset of about two billion tweets was parsed using an algorithm written in Python that looked for specific, well thought-out keywords [2] related to the ten largest stadiums under construction in the United States. *Figure 1* lists the ten American stadiums of focus in the order of projected capacity and provides other important qualities. Unless stated otherwise, affiliated organizations and sports are assumed to be professional.

Figure 1. *Stadium Traits*

City	State	Name	Main Use	Other Uses?	Team(s)	Completion Year	Capacity
Inglewood	CA	SoFi Stadium	Football	Yes	Los Angeles Rams, Los Angeles Chargers	2020	70,000
Las Vegas	NV	Allegiant Stadium	Football	Yes	Los Vegas Raiders, UNLV Rebels	2020	65,000
Birmingham	AL	Protective Stadium	Football	Yes	UAB Blazers	2021	45,000
Arlington	TX	**Globe Life Field**	Baseball	Yes	Texas Rangers	**2020**	40,000
Oakland	CA	Oakland Ballpark	Baseball	Yes	Oakland Athletics	2023	35,000
San Diego	CA	SDSU West Stadium	Football (college)	Yes	San Diego State Aztecs	TBD	35,000
Honolulu	HI	New Aloha Stadium	Football (college)	Yes	University of Hawaii Vili	2023	30,000
Nashville	TN	Nashville Fairgrounds Stadium	Soccer	Yes	Nashville FC	2022	27,500
Cincinnati	OH	West End Stadium	Soccer	Yes	Cincincati FC	2021	26,000
Miami	FL	Miami Freedom Park	Soccer	Yes	Inter Miami CF	2022	25,000

Our various keyword lists related abbreviated versions of the team's names [3], official handles for the organization(s) tied to the stadium, team and stadium names, appropriate slogans, and common hashtags found on twitter when searching for these stadiums on the app. The keywords and text were modified to guarantee matches when applicable [4]. Duplicate data was then handled [5].

While each individual tweet in the original dataset contained over twenty attributes, the dataset this project uses focuses on text [6], location, and time of creation, and language. After filtering the entire dataset[7] and cleaning of repeat and non English data, we found 274, 607 tweets that matched our keywords. Notably, the dataset is further broken down and analysed in a later section of this project.

After gathering the data, sentiment analysis [2] was performed using the Python library *"TextBlob"*. The sentiment analysis reads the text of each tweet and returns the 'sentiment', defined as "a named tuple of the form sentiment(polarity, subjectivity)." [8]

Polarity ranges from [-1.00, 1.00], where (-1) means the text speaks very negatively and (+1) means the text speaks very positively. Subjectivity ranges from [0.00, 1.00], where (0) means the text displays no subjectivity and (+1) means the text is entirely subjective. Proceeding onto the findings from our data, we will first look at the composition of the dataset by stadium.

General Findings

Figure 2. *Percent of Tweets by Stadium*

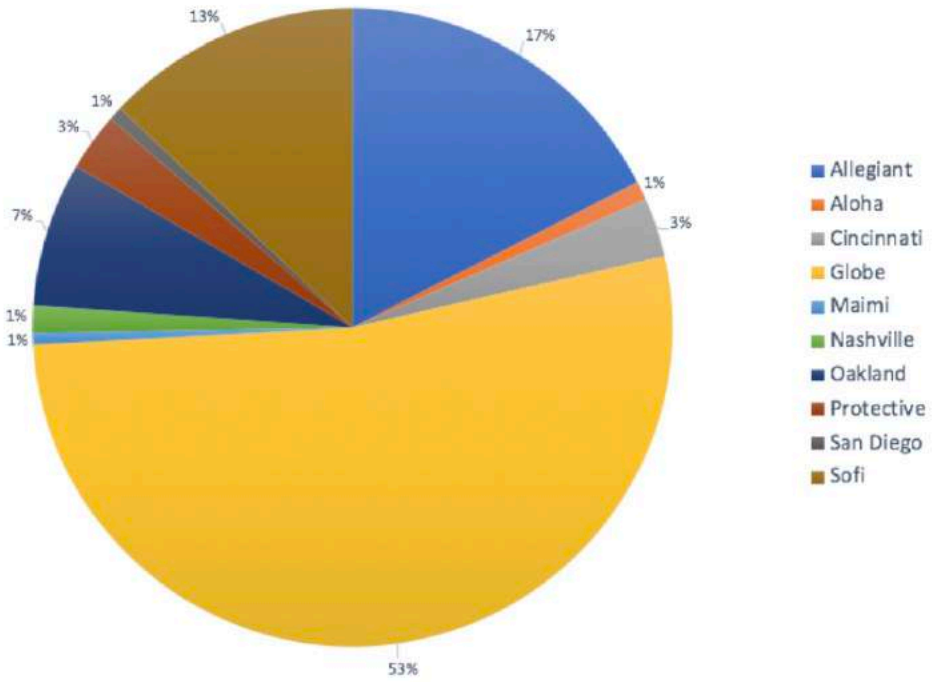


Figure 2 visually breaks down the composition of our data. This figure is a result of organizing the data by which stadium's keyword picked up a specific tweet.

The largest three portions of our dataset came from upcoming Globe Stadium in Arlington, Texas (~53%), Allegiant Stadium in Las Vegas, Nevada (~17%) and Sofi Stadium in Inglewood, California (~13%). Given that these stadiums are three of the four largest and nearest to completion, it is understandable that these stadiums were the most tweeted about.

Throughout the remainder of the project, all figures referring to Polarity specifically will be blue and figures referring to Subjectivity specifically will be red using ten bins with a range of (0.20) for Polarity and ten bins with a range of (0.10) for Subjectivity.

Figure 3A. *Count of Polarity in each Range*

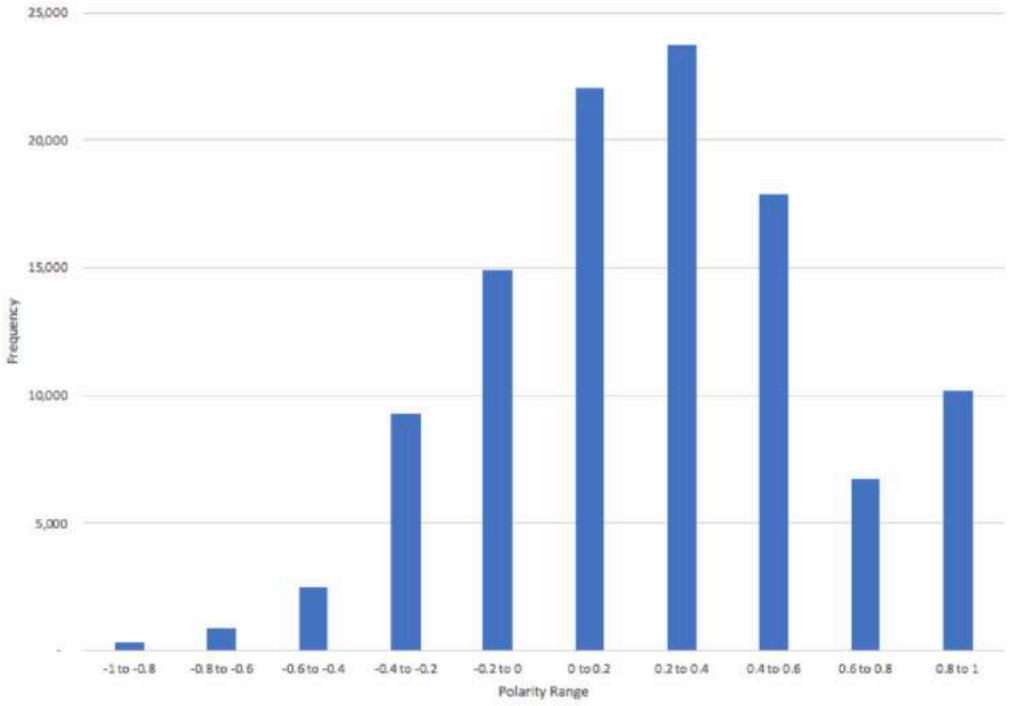
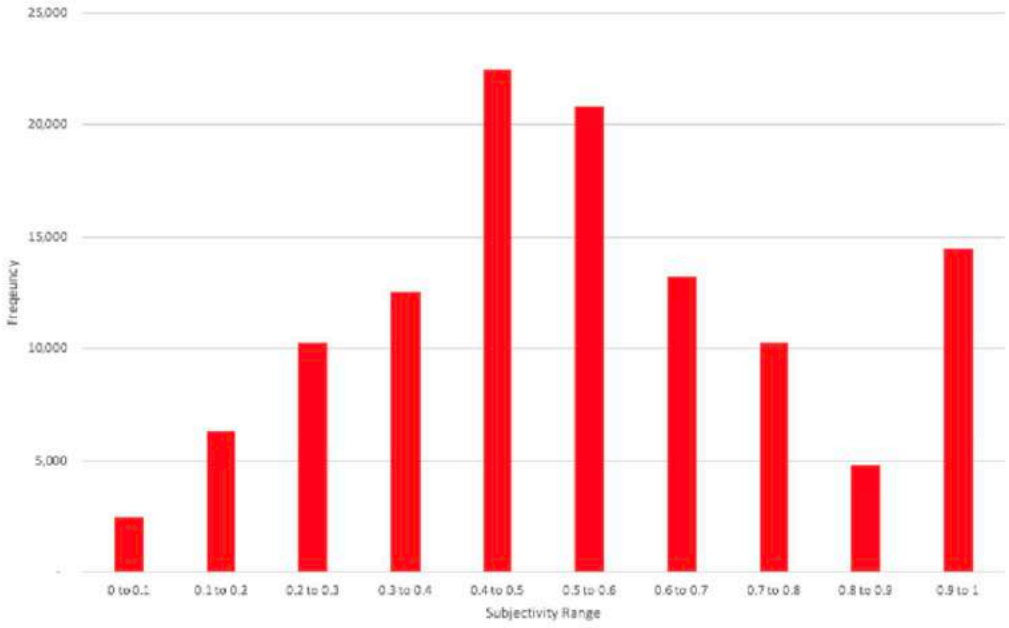


Figure 3B. *Count of Subjectivity in each Range*



On both graphs, the frequency of tweets spikes around the middle of their respective ranges. The polarity appears to be slightly positive, with the majority of tweets having a polarity equal to or greater than zero. Overall, the graph of Subjectivity's average is around 0.5, but there is a spike around (+1.00).

The original dataset spans a time period of roughly twenty five months. Figure 4 portrays average sentiment over time, by splitting this time period into groups of three (knowing the data from the last bin is incomplete).

Figure 4A. *Average Polarity for each Time Period*

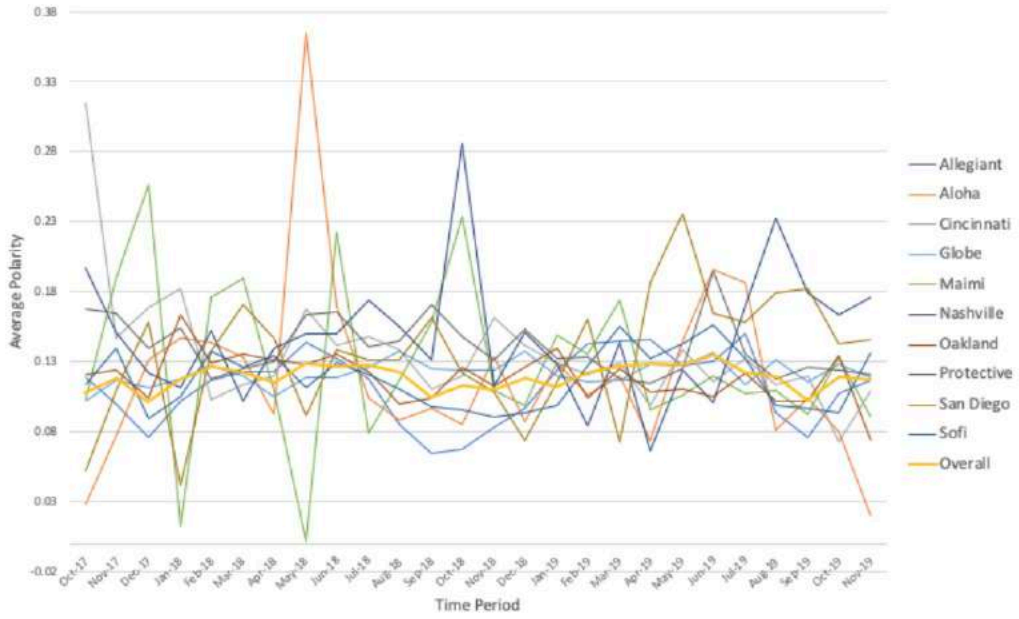
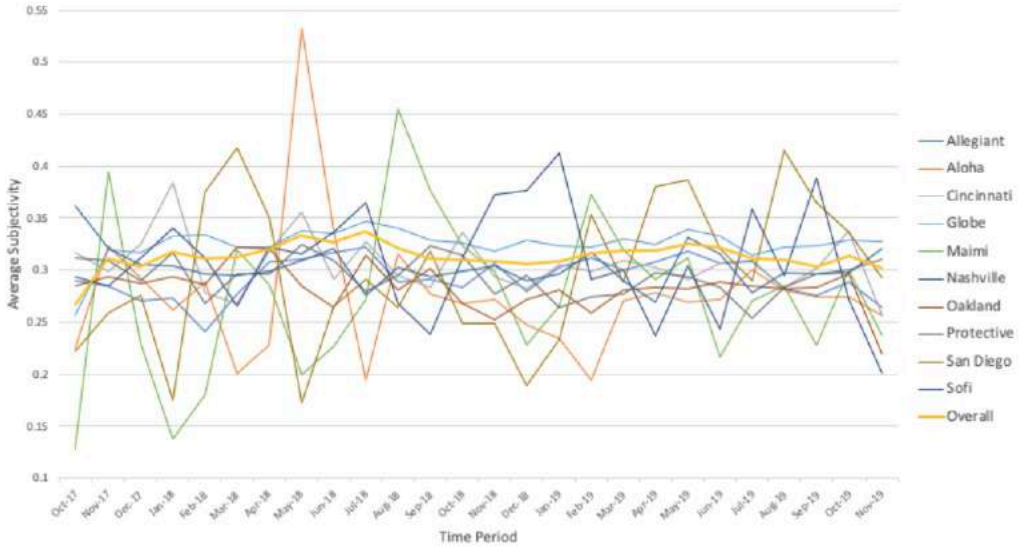


Figure 4B. *Average Subjectivity for each Time Period*



These figures represent the change in polarity and subjectivity over time. While the graphs look fairly hectic, the scale of the Y-axis is very important, as it tells us that there is actually fairly small deviations in both polarity and sentiment. Our polarity data shows that average sentiment of the data was consistently higher than (0.00), at about (0.10), which agrees with the results of Figure 3A. Our subjectivity data shares a similar trend, with the average being slightly higher than (0.30) for most periods. This again matches the expectations from Figure 3B.

Because this project focuses on American stadiums, we used the location of tweets that contained this field to map the tweets by stadium on North America. Below are these maps. [9] [10]

Figure 5A. *Map of Globe Life Data in North America*

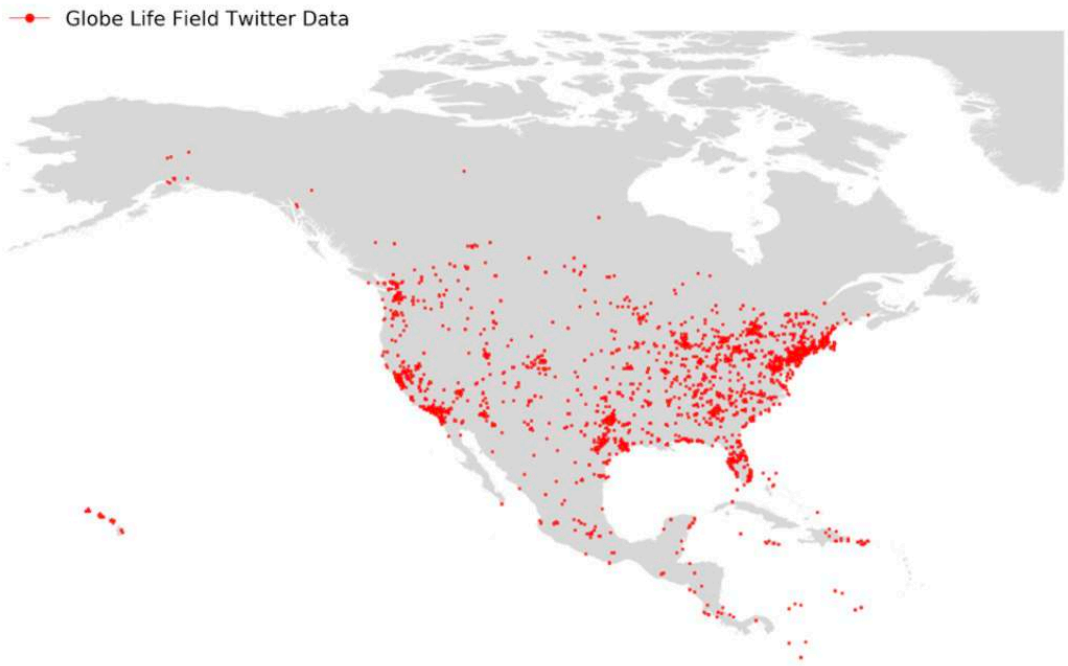
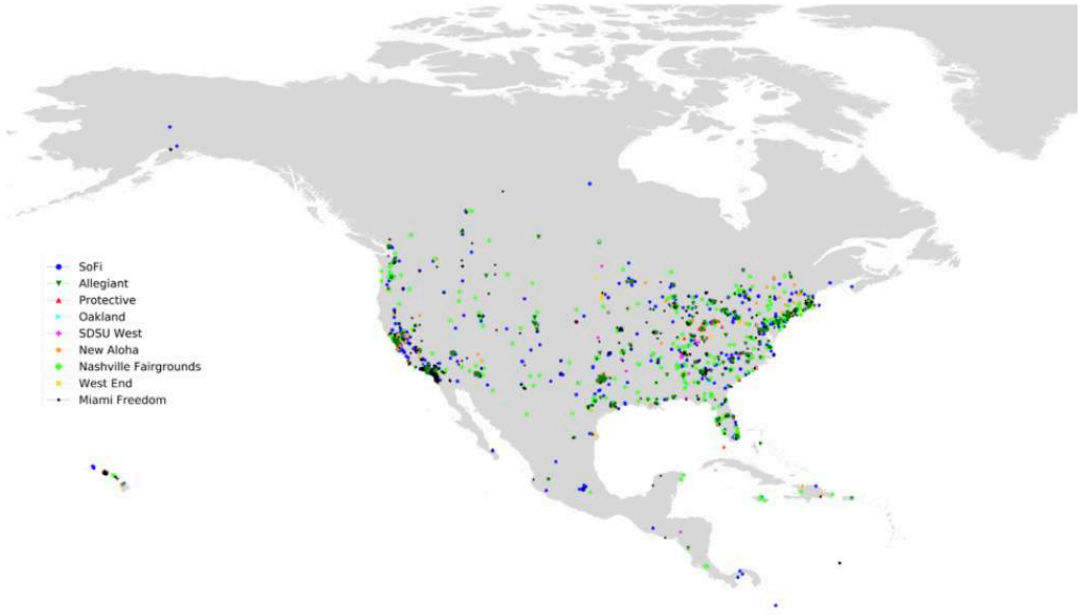


Figure 5B. *Map of all the Other Data in North America*



It is evident in both graphs that the majority of the tweets centered in North America are sent from within the United States, with many hotspots in the Eastern half and West coast. Each of the stadiums also appear to have clusters close to their home city, but in every case, tweets relating to every stadium are found across the nation.

This interesting note most likely comes from our first source of error: our keywords. Because these determined our specific dataset, having keywords pick up tweets that do not talk about the topic at hand is a very big issue. which ultimately determined the tweets to be included in the dataset. Some of our keywords did not refer to the stadiums specifically, but to the teams that played in their respective stadiums. As a result some of our data did not apply to the topic, but was very hard to filter out well. This increased our *Bayes error*[11] drastically.

As mentioned earlier, there is about as much data for Globe Life Field as the rest of the nine stadiums combined. Because Globe Life represented so much of the total dataset, we decided to focus the rest of our efforts on Globe Life's data specifically. Selecting one stadium also simplified the training of our models[12], which is discussed later on. We assumed that because of the above factors, it would be a decent proxy for many of the other stadiums.

Globe Life Field's Data Findings

Figure 6A. *Globe Life Field's Average Polarity over Time*

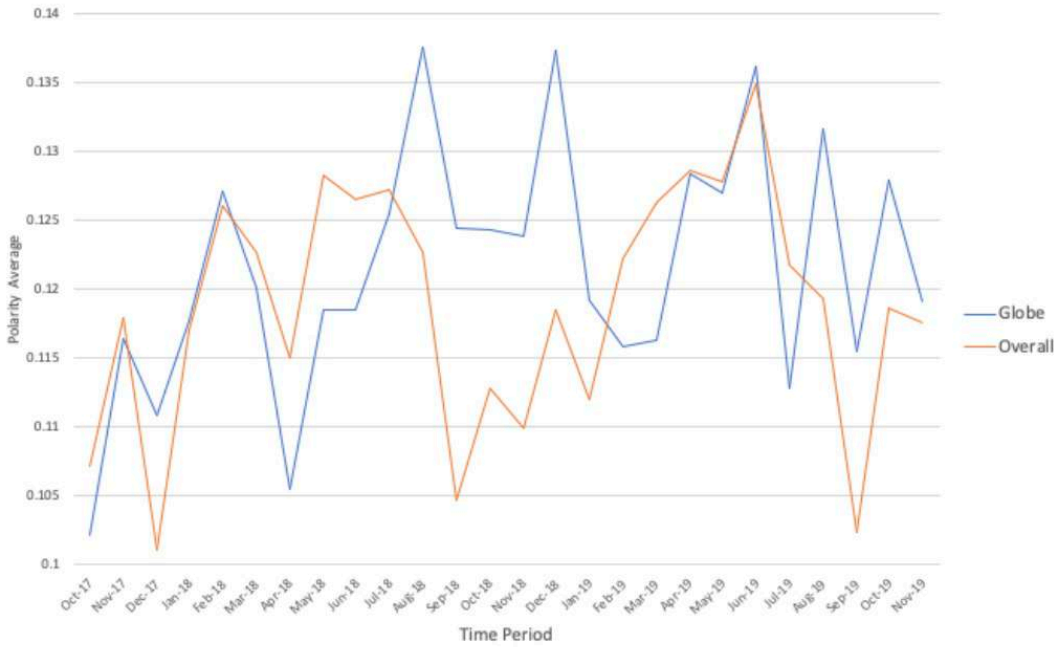
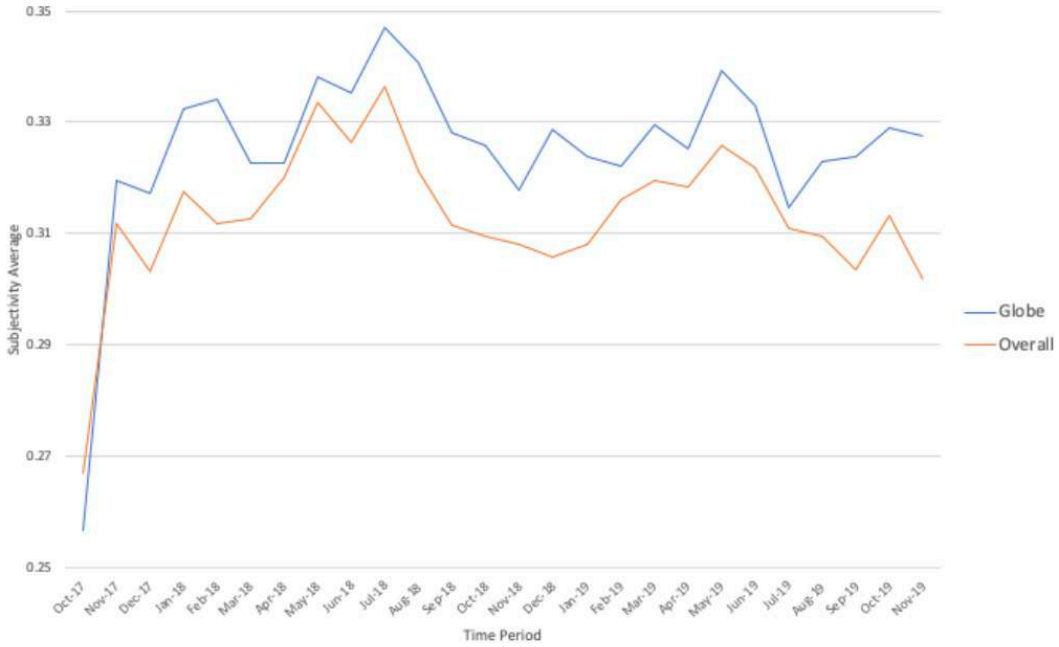


Figure 6B. *Globe Life Field's Average Subjectivity over Time*



Over the same nine time ranges, Globe Life's polarity appears to be very similar to that of the entire dataset, reinforcing our choice for the single stadium to represent our entire dataset. Over this entire time period, the average of the Globe Life data deviated a maximum of just over (0.03) from the overall dataset, suggesting that it was a very good approximation for the data as a whole. Our subjectivities deviation is even smaller. The trend of the Globe life data and of the entire dataset appear to move in sync over all twenty six months.

Conclusions

As we have discovered, predicting the sentiment towards the construction of stadiums is a challenging problem. However, we found that the average sentiment towards a stadium as construction nears completion does not change drastically. While months until completion alone was an extremely poor predictor of either sentiment or polarity, using the other term (sentiment for polarity and polarity for sentiment) gave a better answer. However, because polarity and sentiment are not known without the other, the best approximation of polarity or sentiment at a given time would be a range that is dependent on a range of either sentiments or polarities.

After optimizations of OLS[13], ridge and lasso linear models on this dataset, OLS was found to be the best of the three linear models. However, we still consistently had very mean squared error, meaning that the model was a poor predictor of the sentiment at any given point in the data.

This poor fit comes from our extremely large Bayes error that comes from the high variation of the data within the dataset. As a result, there is a very high amount of irreducible error within this dataset. However, if we were to attempt this or a similar project again, this could be minimized by choosing keywords that acted as a better[14] filter on the general dataset and therefore would give us a dataset with improved relation to our variables of interest. On the other hand, our estimation error was extremely low, as our dataset was so large that our model was a good approximation for the true value of the entire population.

Technical Appendix

Throughout the project, the data was saved in CSV files that were accessible using Python script or Microsoft Excel. All of the summary figures for the data in the "General Findings" and "Globe Life Field's Data Findings" sections were made using Excel. After this section, every other figure was generated using Python script on the same CSV files. For all models, the data was broken up into a training set and a test set, where the training set consisted of 13,800 data points and the test set consisted of the other 9,850 data points.

For our basic one-dimensional linear regression, we performed this regression in two steps. In the first, we used Pythons Pandas library followed by SkLearn's linear regression to build the model. Because it was a very simple model, there was no optimization done. Next, we used the Seaborn library to construct a visual of this model and regression visuals.

For the OLS, ridge and lasso models with multiple independent variables, we used SkLearn's linear model. First, to visualize the three-dimensional data, we used Python's MatPlotLib's feature of "mplot3d." Since MatPlotLib is only intended for use with two dimensional graphs, the functions relating to three dimensional graphs are very limited. As a result, these graphs were not nearly as well designed as we would have liked, but we did not have the functionality to improve them. OLS could not be optimized, so we did not modify this command at all. Because ridge regression takes an 'alpha' term as a parameter, we optimized this alpha term to minimize the mean squared error of the model on the test data. After optimizing alpha, a ridge model was trained on the data. Finally, we performed this same method for the lasso models that performed poorly.

After performing all of these regressions, SkLearn's 'metrics' library was used to find the R-squared values and the Root Mean Squared Error terms for every model. These were then saved, transferred to an Excel document and compared against each other.

Footnotes

1. Public funding is justified since the organization brings their brand and is supposed to increase economic activity around the area. For more on the history of athletic facilities and their role in America see Johnson Garrett [1] , "The Economic Impact of New Stadiums and Arenas on Cities", *University of Denver Sports & Entertainment Law Journal 10*, (2011)
2. **Keyword Lists** Here are the lists of keywords by stadium used to filter tweets from the original dataset:

   Sofi:

   sofistadium, sofi stadium, nfl la, la football, ramsNFL,RamsHouse, LARams, la rams, rams fan, @Chargers, BoltUp, LAChargers, la chargers, and chargers fan.

   Allegiant:

   AllegiantStadm, allegiantstadium, allegiant stadium, Las Vegas Raiders, Lv raiders, Oakland raiders, @Raiders, #Raiders, RaiderNation, raider nation, raider fans, and raiders stadium.

   Protective:

   Protectivestadium, protective stadium, protectivestdm, University of Alabama at Burmingham Football, UAB football, go blazers, goblazers, UABgreengang, UAB_FB, and BJCC.

   Globe:

   Globelifefield, globe life field, globe field, globefield, rangers fan, rangers ballpark,rangers stadium, rangers field, rangersballpark, rangersstadium, and rangersfield.

   Oakland:

   oakstadium, oak stadium, oaklandstadium, oakland stadium, oaklandathletics, oakland athletics, @athletics, bjarke ingels, howard terminal, howardterminal, and jack london square.

   SDSU West:

   san diego west stadium, SDSUWest, sd west stadium, san diego stadium, sandiegostadium, Weststadium, sdweststadium, san diego football, sd football, sdfootball, mission valley Stadium, sdsu mission valley, and sdccu stadium.

   New Aloha:

   new aloha stadium, aloha stadium, newaloha, Aloha Stadium Hawaii, hawaiifb, AlohaStadiumHI, newalohastadium, aloha stadium, hawaii football, hawaiifootball, u of hawaii football, and hawaiifb.

   Nashville Fairgrounds:

   nashville fairgrouds, FAIRGROUNDSNASH, nashvillefairgrounds, nashville fairground, nashvillefairgound, fc nashville, fcnashville, nashville soccer, nashvillesoccer, nashville sc, and nashvillesc.

   Cincinnati:

   cincinnati stadium, cincinnati stadium, cincinati stadium, FCCincy, cincinatistadium, west end stadium, westendstadium, westendstdm,mcincinnati fc,mcincinnatifc, cincinatifc, fc cincinnati,fccincinnati,fc cincinati, and fccincinati.

   Miami Freedom:

   miami freedom stadium, miafreedompark, miamifreedomstadium, miami stadium, Inter Miami CF stadium, miamistadium, inter miami, inter maimi, miami cf, miami fc, intermiami, Intermaimi, miamicf, miamifc, intermiami, intermaimi, InterMiamiCF, freedom stadium, and freedomstadium.
3. The keyword 'larams' could be found in the text of tweets mentioning '@larams' , '#larams, 'larams.', etc.

4. Keywords (Python strings) were uppercased and checked whether they were contained in the tweet's uppercased text (also strings). While parsing the original dataset, this eliminated the need to align casing between keywords and potential matches within the tweet's text. The documentation for Python's string library provides methods to handle casing. With this method, there was a possibility of multiple keywords appearing in a single tweet, creating duplicates.
5. Duplicate data was found and deleted by searching the entire dataset of filtered tweets for instances of tweets where the text was the same. If a tweet's text was found to be the same, all instances after the first one were deleted.
6. Commonly found in actual tweets are emojis, pictures, and/or gifs and these content formats were not taken into consideration during the sentiment analysis.
7. approximately two billion tweets
8. The Textblob library provides access to simple natural language processing tasks such as sentiment analysis.
9. We used GeoPandas, an open source project, to make working with geolocation data in Python easier. The coordinate data was plotted onto North America (using this shape-file).
10. About 1% of users enable geolocation, so the below maps only show the location of tweets with location enabled in North America. Because this 1% is consistent among all users, it is a decent proxy for the distribution of our entire dataset.
11. The best (lowest) possible error rate
12. Linear models trained to predict the sentiment of a particular stadium given months away from completion
13. ordinary least squares
14. Aside from the fact that this tweet is recent (beyond the scope of the original twitter data set, no keyword from any list would have picked up an opinionated tweet such as this one. The sentiment analysis results from *TextBlob* for this particular tweet are:
    Polarity = 0.487273 are Subjectvitity = 0.570909

References

1. **The Economic Impact of New Stadiums and Arenas on Cities**
   Johnson, G., 2011. University of Denver Sports & Entertainment Law Journal, 10.
2. **textblob Documentation**
   Loria, S., 2018.

Citations and Reuse

For attribution in academic contexts, please cite this work as

```
Novitsky & Patel, "Don’t expect the public to look at your stadium any different as you come
```

BibTeX citation

```
@article{novitsky2019don’t,
  author = {Novitsky, Sasha and Patel, Afzal},
  title = {Don’t expect the public to look at your stadium any different as you come close to
  journal = {Data Insights},
  year = {2019}
}
```