

Content Analysis of Tweets Using NHL Team Specific Hashtags

Carly Rose Willing

18 December 2017

1 Introduction

Social media usage has infiltrated most aspects of everyday life. Most people have one, if not more, social media accounts, and use these social networks, in some capacity, each and every day. In fact, Statista estimates that there are 2.46 billion social media users, and Twitter has an average of 330 million monthly users. Tweets, previously limited to only 140 characters, can reveal a lot about an individual, and their perceptions of current events, brands, and other aspects of daily life, which are becoming increasingly public.

1.1 Why Hockey?

When presented with this project, the idea of evaluating the widespread perceptions of brands seemed dull. However, the night after the project was assigned, I found myself lurking Twitter for updates regarding upcoming trades by my favorite NHL teams. I realized that each of the 31 NHL teams all Tweeted using a hashtag specific to their team. Whenever someone uses one of these specific hashtags, it appears with the team logo next to it. Ultimately, I decided that I could use Twitter to analyze the perceptions of each team, and possibly use Twitter data to predict how well or how poorly a team was doing at any given time.

Today's #NHL Games:



















#GoKingsGo 	vs	#NJDevils 
#Sens 	vs	#Sabres 
#TMLtalk 	vs	#LetsGoFlyers 
#GoAvsGo 	vs	#ALLCAPS 
#LetsGoOilers 	vs	#CBJ 
#GoBolts 	vs	#AllTogetherNowSTL 
#CofRed 	vs	#MNWild 
#FlaPanthers 	vs	#Blackhawks 
#Redvolution 	vs	#VegasBorn 

Figure 1: Various Team Hashtags

1.2 The NHL and Twitter

After looking at each individual teams Twitter feed, I realized that each team had their own specific way of interacting with fans. For example, the Vegas Golden Knights Twitter is full of sarcastic jokes, humor, and memes (Figure 2). I also noted that teams that were doing worse in terms of overall ranking in their division, seemed to publish Tweets that were less enthusiastic than teams that were doing better. As one would assume, teams who Tweeted more often and engaged more frequently with fans had a higher number of followers, and a higher rate of use of their team-specific hashtag. I also noticed that most of the Tweets using the team-specific hashtag that were not published by the team themselves, were most often used by fans of the team. Thus, each Tweet had a very straight forward, unfiltered perception of the team. As I was reading Tweets from my peronally favorite team, the Chicago Blackhawks, I noticed that many people had optimistic and positive reactions to an increase in goals scored over the last few games, but following a miserable loss, the Tweets reflected the sentiment. Table 1 shows the name of each team, their corresponding hashtag, follower count (rounded to the nearest thousand), and the division they are in.



Figure 2: Vegas Golden Knights Twitter

Table 1: NHL Twitter Presence

Teamname	Official Hashtag	# Followers	Division
Anaheim Ducks	#LetsGoDucks	556,000	Pacific
Arizona Coyotes	#Yotes	329,000	Pacific
Boston Bruins	#NHLBruins	1,340,000	Atlantic
Buffalo Sabres	#Sabres	638,000	Atlantic
Calgary Flames	#CofRed	584,000	Pacific
Carolina Hurricanes	#Redvolution	335,000	Metropolitan
Chicago Blackhawks	#Blackhawks	2,450,000	Central
Colorado Avalanche	#GoAvsGo	423,000	Central
Columbus Blue Jackets	#CBJ	418,000	Metropolitan
Dallas Stars	#GoStars	932,000	Central
Detroit Red Wings	#LGRW	1,210,000	Atlantic
Edmonton Oilers	#LetsGoOilers	781,000	Pacific
Florida Panthers	#FlaPanthers	340,000	Atlantic
Los Angeles Kings	#LAKings	1,150,000	Pacific
Minnesota Wild	#MNWild	649,000	Central
Montreal Canadiens	#GoHabsGo	1,500,000	Atlantic
Nashville Predators	#Preds	518,000	Central
New Jersey Devils	#NJDevils	679,000	Metropolitan
New York Islanders	#Isles	477,000	Metropolitan
New York Rangers	#NYR	1,380,000	Metropolitan
Ottawa Senators	#Sens	551,000	Atlantic
Philadelphia Flyers	#LetsGoFlyers	1,430,000	Metropolitan
Pittsburgh Penguins	#LetsGoPens	1,690,000	Metropolitan
San Jose Sharks	#SJSharks	774,000	Pacific
St. Louis Blues	#AllTogetherNowSTL	633,000	Central
Tampa Bay Lightning	#GoBolts	634,000	Atlantic
Toronto Maple Leafs	#TMLTalk	1,740,000	Atlantic
Vancouver Canucks	#Canucks	1,020,000	Pacific
Vegas Golden Knights	#VegasBorn	261,000	Pacific
Washington Capitals	#ALLCAPS	667,000	Metropolitan
Winnipeg Jets	#GoJetsGo	492,000	Central

2 Mining the Tweets

I collected 500 Tweets from each of the 31 team-specific hashtags, for a total of 15,500 Tweets, and filtered out all retweets in order to avoid unnecessary repetition of text. The Tweets were all collected from the period October 04 2017 (the official start of the NHL seasons), and December 15 2017. The Tweets were cleaned of all usernames, hyperlinks, punctuation, and unnecessary blank characters or emojis. I found that the current standings of the teams,

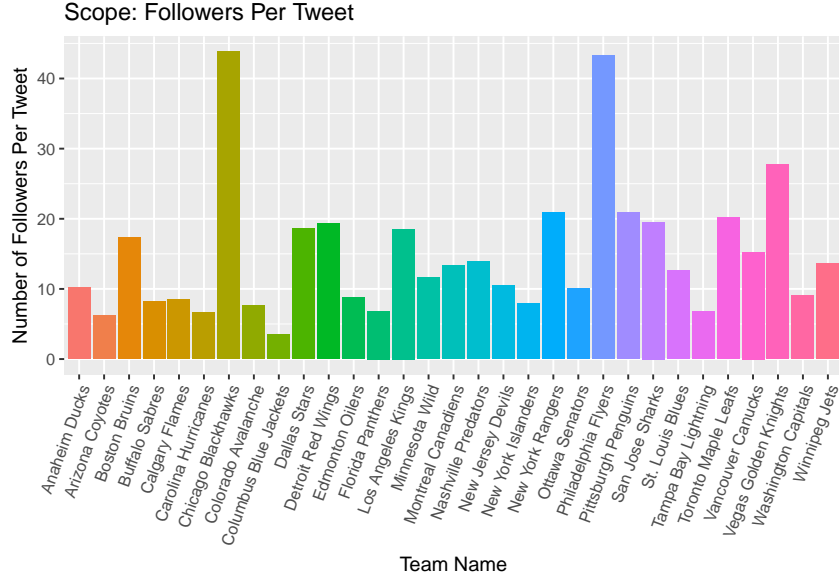


Figure 3: Scope of Twitter Reach

as opposed to performance early on in the season, was the best classification method. In order to look at the difference in content between teams doing well, and teams doing poorly, I separated the 31 teams into three unique groups– “good” teams, “bad” teams, and “neutral” teams. The good teams are comprised of the top two teams in each of the four divisions. The bad teams are comprised of the bottom two teams in each of the four divisions. Finally, the neutral teams are those that fall in the middle of their division. After the Tweets were separated into datasets following the below separation, using a dictionary of stopwords including traditional English stopwords, the search terms used in the initial data mining, and corresponding team and location names, the Tweets were unnested into a data structure with each individual word.

2.1 Team Classification

“Good” Teams

- Tampa Bay Lightning
- Nashville Predators
- Los Angeles Kings
- Toronto Maple Leafs
- Columbus Blue Jackets
- Vegas Golden Knights
- St. Louis Blues
- Washington Capitals

“Bad” Teams

- Ottawa Senators
- Buffalo Sabres
- Philadelphia Flyers
- Colorado Avalanche
- Arizona Coyotes
- Chicago Blackhawks
- Edmonton Oilers
- Carolina Hurricanes

“Neutral” Teams

- Montreal Canadiens
- Boston Bruins
- Detroit Red Wings
- New York Islanders
- Pittsburgh Penguins
- Dallas Stars
- Minnesota Wild
- Vancouver Canucks
- San Jose Sharks
- Calgary Flames
- Anaheim Ducks
- New York Rangers
- Florida Panthers

2.2 Visualizing Frequency of Terms

Using the unnested data sets, I created wordclouds for both the good dataset and the bad dataset, which showed the frequency of word usage. Interestingly enough, one of the top used words by “Bad” teams was actually the word “win”.



Figure 4: Word Frequency for Good Teams



Figure 5: Word Frequency for Bad Teams

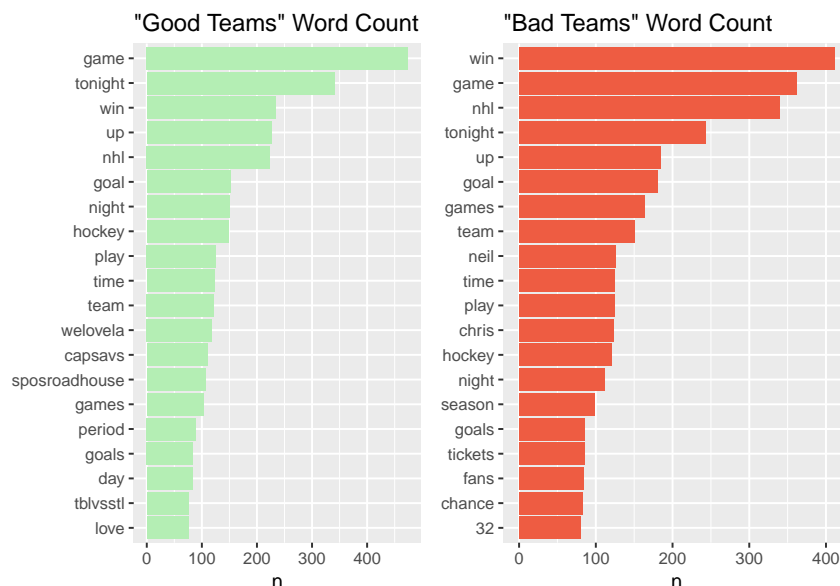


Figure 6: Comparative Word Counts

3 Sentiment Analysis

3.1 Sentiment Scores of Words

After looking at frequency of words used in Tweets related to good teams or bad teams, I used a custom sentiment dictionary which included a pre-built dictionary, using the affinn format, as well as my personal addition of hockey related terms, to score each of the words with a corresponding numeric value from -5 to 5. -5 corresponds to the terms which are seen as the most negative, while terms with a score of 5 are the most positive. It can be noted that there are no -5 scored words in the comparison cloud for good teams, but there is both -5 and 5 scored words for the bad teams. Analysis on a team by team basis can be seen in the corresponding Shiny application.

3.2 Team Name vs. Sentiment

It can be seen in the corresponding chart that each team had a varying average sentiment score, calculated by the sum of the scores, divided by the total number of words. An outlier, as seen in the bar of the Columbus Blue Jackets is seen, as the average sentiment score is significantly lower than the other teams. While one would think that Columbus would have a high average sentiment do to their winning record this season, and being second in their division, the team has recently been plagued with numerous injuries affecting some of their more seasoned key players. Further evaluation of the sentiment analysis shows that there are many terms mentioning the orbital fracture sustained by forward Brandon Dubinsky in a fight last week, putting him on the injured list for the next six to eight weeks. After looking at the correlation between the number of points a team has (a sum of twice their number



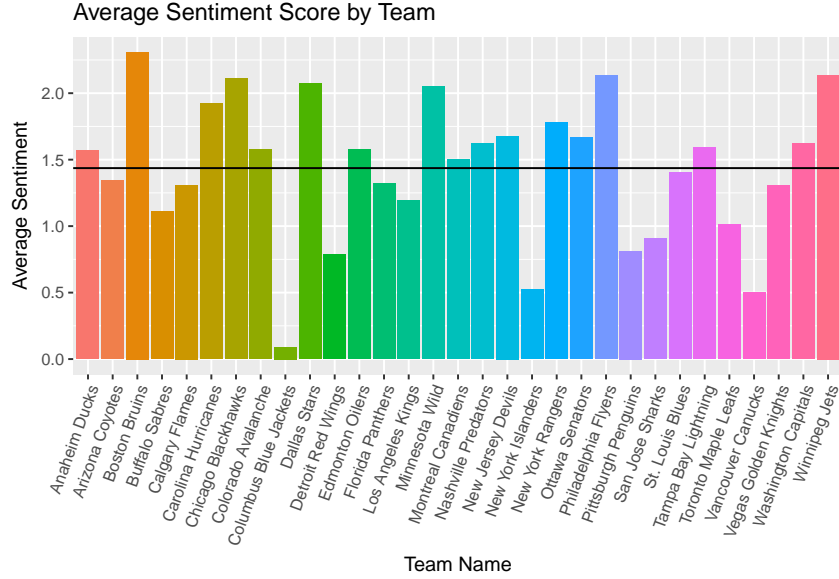


Figure 9: Comparison of Average Sentiment Scores by Team

of wins plus the number of overtime losses), and their average sentiment score, there was only a correlation of 0.02510923. This indicates that there is no relationship between the two variables. However, interaction with other predictor variables, including the number of Twitter followers a team has, may prove valuable in model fitting in terms of prediction of points. A summary table of the average sentiments is found below. In order to evaluate team performance further, I employed the use of the weekly power ranking as released by ESPN. The power ranking takes into account improvement or decline within the last week, and has a 0.95 correlation to the number of points a team has, making it a good unbiased estimator of pervasive team performance.

3.3 Sentiment vs. Average Length of Characters

After evaluating some of the Tweets, I noticed that longer Tweets tended to be indicative of rather negative sentiments, often long rants regarding the playing style of a particular player. While the scatterplot indicates that there may be some evidence of a correlation, mathematical analysis concludes that there is a -0.1015794 correlation between the two, indicating a lack of significant relationship.

4 Mapping Sentiments

In order to better visualize the distribution of sentiments from team to team, the following plots were created. These plots do a good job of demonstrating to what extent each team is associated with different sentiment scores. Figure 12 shows that the use of terms with sentiments of -5 are limited to only three teams, while most teams across the nation had a

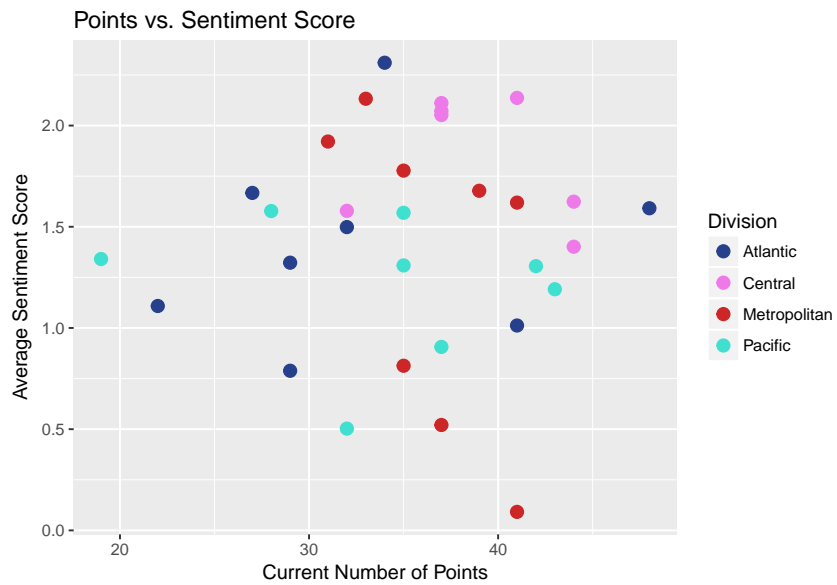


Figure 10: Scatterplot of correlation between average sentiment score and current number of points

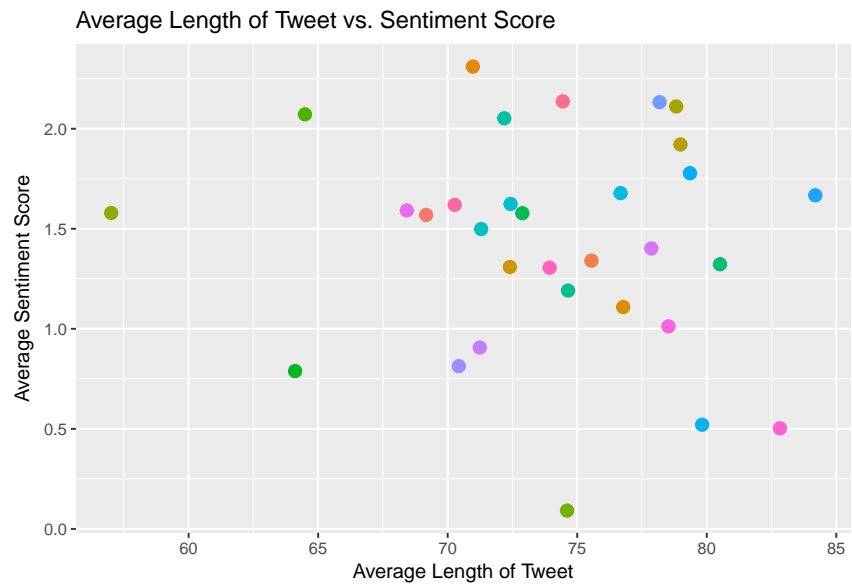


Figure 11: Sentiment score compared to average length of Tweet

Table 2: NHL Team Rankings and Sentiments

Teamname	Power Ranking	Points	Classifier	Sentiment Score
Anaheim Ducks	19	35	Neutral	1.56896552
Arizona Coyotes	31	19	Bad	1.34068627
Boston Bruins	16	34	Neutral	2.31043257
Buffalo Sabres	30	22	Bad	1.10858586
Calgary Flames	20	35	Neutral	1.30900243
Carolina Hurricanes	23	31	Bad	1.92111369
Chicago Blackhawks	17	37	Bad	2.11111111
Colorado Avalanche	25	32	Bad	1.57881137
Columbus Blue Jackets	5	41	Good	0.09144543
Dallas Stars	15	37	Neutral	2.07178218
Detroit Red Wings	28	29	Neutral	0.78861789
Edmonton Oilers	26	28	Bad	1.57772021
Florida Panthers	27	29	Neutral	1.32250000
Los Angeles Kings	6	43	Good	1.19128329
Minnesota Wild	12	37	Neutral	2.05206074
Montreal Canadiens	24	32	Neutral	1.49839228
Nashville Predators	3	44	Good	1.62407862
New Jersey Devils	9	39	Neutral	1.67796610
New York Islanders	10	37	Neutral	0.52073733
New York Rangers	13	35	Neutral	1.77747253
Ottawa Senators	29	27	Bad	1.66727273
Philadelphia Flyers	21	33	Bad	2.13255814
Pittsburgh Penguins	18	35	Neutral	0.81341108
San Jose Sharks	14	37	Neutral	0.90641711
St. Louis Blues	2	44	Good	1.40144231
Tampa Bay Lightning	1	48	Good	1.59148936
Toronto Maple Leafs	7	41	Good	1.01219512
Vancouver Canucks	22	32	Neutral	0.50279330
Vegas Golden Knights	11	42	Good	1.30548926
Washington Capitals	8	41	Good	1.61942257
Winnipeg Jets	4	41	Neutral	2.13669065

high proportion of terms with sentiments of -2 and -1. Additionally, Figure 15 demonstrates the large proportion of terms of sentiment 5, an encouraging sign for the overall tone of social media.

Distribution of Negative Sentiments

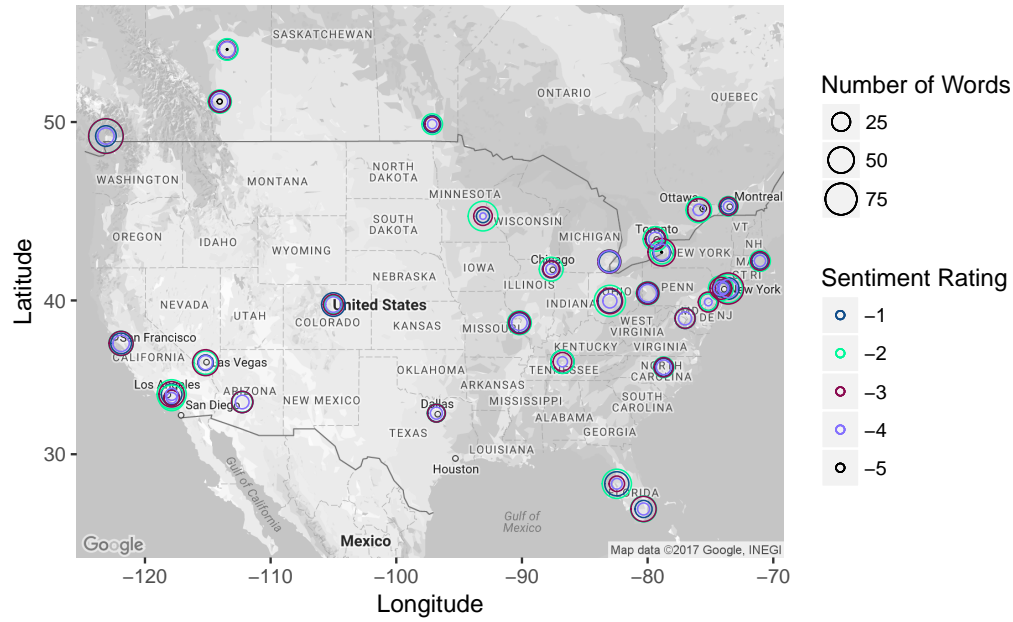


Figure 12: Frequency of negative sentiments used for all teams

Distribution of Negative Sentiments in Tri-City Area

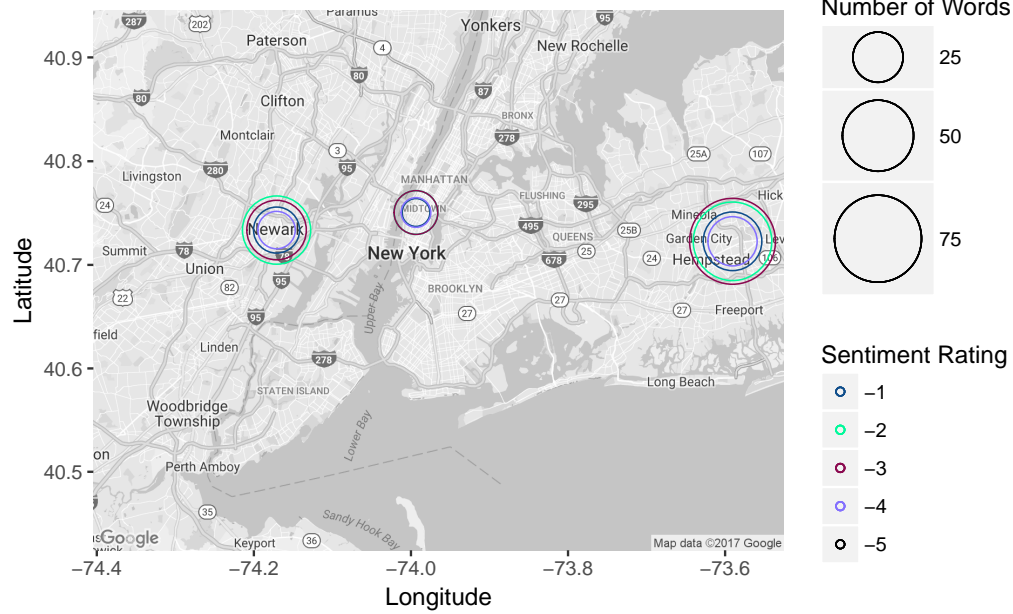


Figure 13: Frequency of negative sentiments used for all teams: zoom on tri-city area

Distribution of Negative Sentiments in Southern California

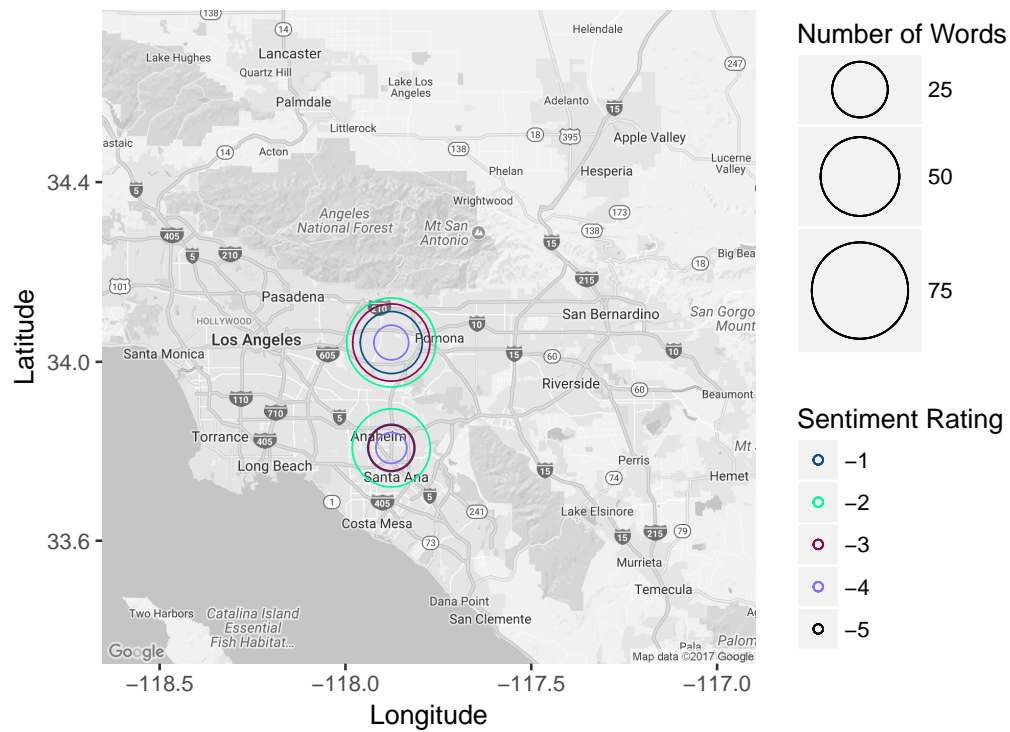


Figure 14: Frequency of negative sentiments used for all teams: zoom on Southern California

Distribution of Positive Sentiments

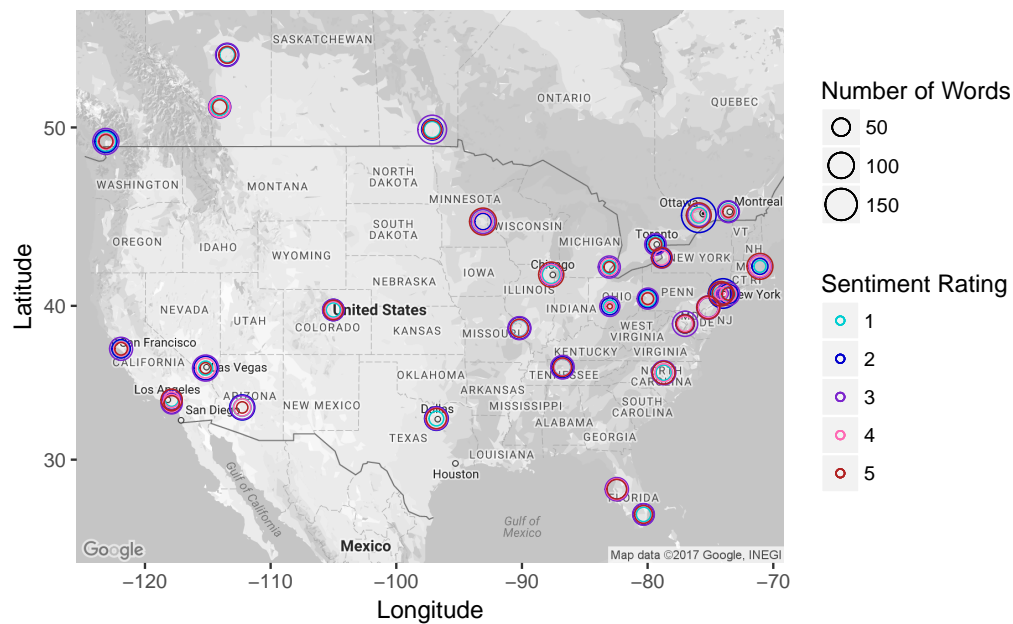


Figure 15: Frequency of positive sentiments used for all teams

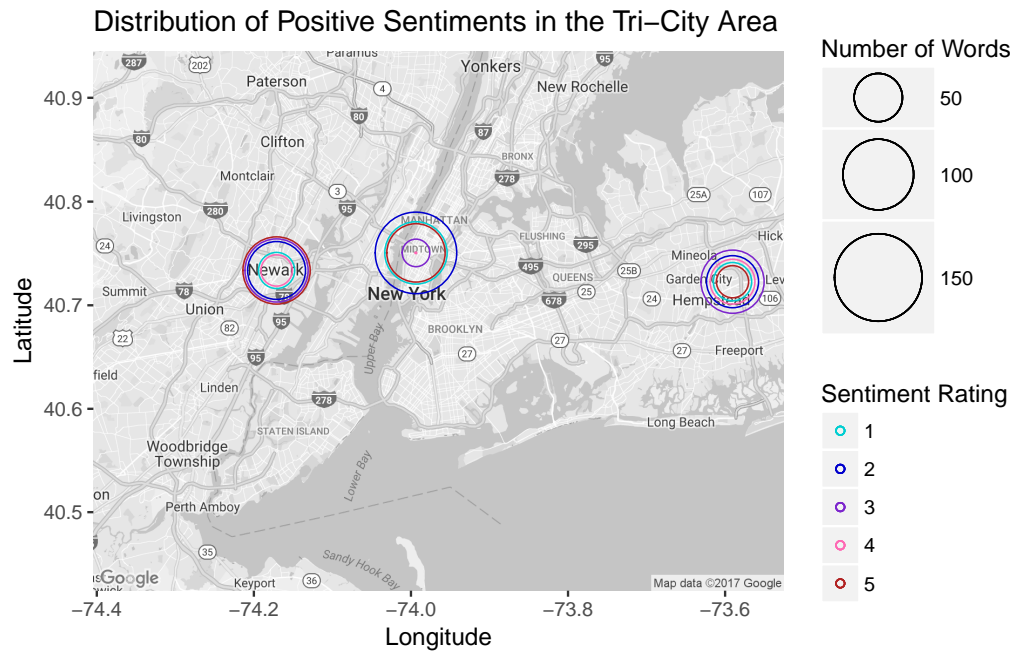


Figure 16: Frequency of positive sentiments used for all teams: zoom on tri-city area

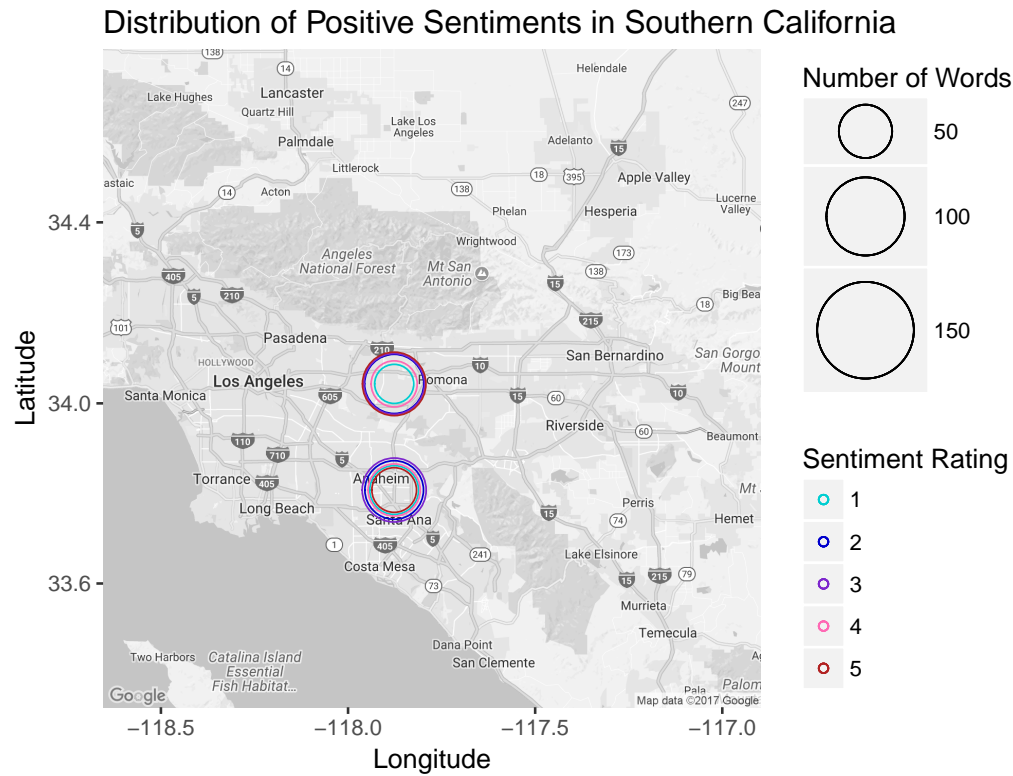


Figure 17: Frequency of positive sentiments used for all teams: zoom on Southern California

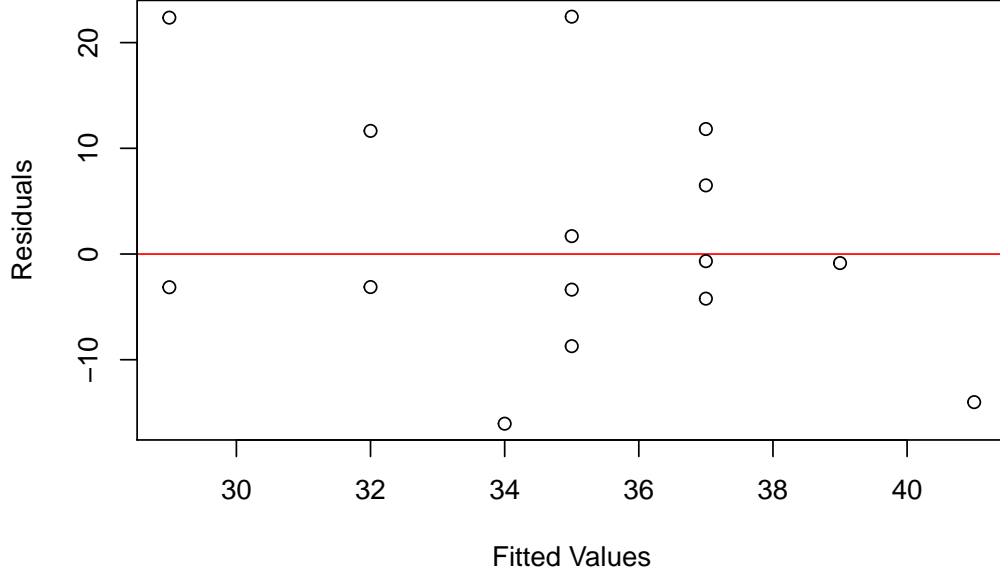


Figure 18: Residual analysis of predictive power of formulated model

5 Linear Model Creation

After running diagnostic plots on the predictor variables, I settled on the following OLS model.

$$\begin{aligned} \text{Points} = & \beta_0 + \beta_1(\text{avgSentimentScore} * \text{totalSentiment Score}) + \\ & \beta_2(\log(\text{FollowersPerTweet})) + \beta_3(\text{LengthofTweets}) + \\ & \beta_4(\text{Division}) + e \end{aligned} \quad (1)$$

I used the Good and Bad teams together to create the model. I then assessed the predictive ability of the model in terms of the Neutral dataset. While some of the residuals are fairly large, they all appear to be randomly distributed. Further tweaking of the model may increase the quality of fit. This was not the case when attempting to predict the Good or Bad data based off the Neutral data, or using only the Good or Bad data to predict the Neutral data. As expected, creating a model based off only the Good data tended to overestimate while the Bad data tended to underestimate.

6 Conclusion

Ultimately, it turns out that due to the variability in a teams performance from game to game, and even period to period of play, sentiment analysis was not indicative of typical measures of team performance. This season may also be an outlier in terms of predictive power due to the large gap between teams within the four divisions. A point differential as

extreme as the one observed between Pacific first place team, LA Kings, and Pacific last place team, Arizona Coyotes, is not typically seen, especially when the season is not even half way over. Further creation of predictive model would benefit from use of multiple seasons of data, or collection and analysis of more Tweets.