Jackie Mares
Misty Peng

LIN 350 Final Report

The NFL is a hobby common among Americans, creating multiple connections across different states. The NFL is shared and valued by through different means of media, whether it be through television reporting or our main focus, which is through various social media platforms where fans of the same team or of other teams express their opinions with the team or a certain team player's performance through social media platforms; Twitter is the social media platform that we derive our information and analysis from. Football and the NFL also give an insight to American culture and how many different social media users will gravitate towards the same trend in their posts. We were motivated to research this subject by the possibility of discovering potential trends with the usage of adjectives, nouns, and all of those culminating in an overall polarity correlating with the results of games played before and after. The NFL holds the interest of millions of Americans and this allows us to analyze and see how different teams and players bring out the positive and negative support of their fans. With our research, we questioned how the polarity of fans' tweets shifts with changes in a team's success as well as changes in this polarity in the expectation of a successful or unsuccessful game. The null hypotheses in this case include "Polarity of tweets the week prior to the a game does not affect the result of the game" and "The results of a match do not affect the polarity of the tweets concerning that match the week after." Our hypotheses included if polarity had a correlation with the success and performance of a team. We also wanted to know if there are any trends that were consistent in the following seasons. The trends found in our research can also play an important part when teams are wanting to strategize during draft season; they want to look at the trends as another way of discovering which player brings out positive opinions from fans (the more positive the opinions are, you can from the trends that the player is valued by the public and that their talent is being noticed by fans). Although Twitter users could not have an official voice in the decisions of NFL draft picks, they are certainly a strong force that can sway opinions, and this research is simply a start of a way to quantify the public's opinion into something the officials in the NFL could confidently use. The research can also conclude to which players bring out a more positive effect on the public or which player brings out a negative response from the public; the same could be concluded by finding a similar trend with a team.

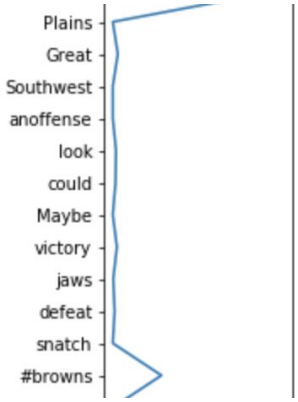
Our tweet data consisted of tweet IDs already compiled by Noah's Ark led by Noah Smith. The data includes six CSV files of approximately one million rows each. The data includes information tweets during every game of the NFL seasons spanning three years from 2010 - 2012. Each year has two CSVs, one being a collection of tweet IDs occurring the week before the game and one being a collection of tweet IDs occurring the week after the game. These were each converted into a data frame, of information include 'tweet_id', 'tweet_UTCtime', 'team', 'opponent', 'week', 'home_away', 'score', 'opponent_score', 'point_spread', 'over_under'. The column 'tweet_id' corresponds with a tweet from the Twitter API, 'tweet_UTCtime' is a record of the time and date at which the tweet was posted, 'team' and 'opponent' are two or three letter abbreviations representing the teams that played in the specified game, 'week' and 'home_away' are further information about which week of the NFL season the game was played as well as where, 'score' and 'opponent_score' show the

corresponding scores, ‘point_spread’ and ‘over_under’ represent how this game’s points compare the team’s overall. For our purposes, we used the columns ‘tweet_id’, ‘team’, ‘opponent’, ‘week’, with additional columns named ‘winning_team’, ‘score_difference’, ‘tweet_text’, ‘tweet_tokenized’, ‘contentwords_only’, ‘polarity_yesno’, and ‘polarity_score’. The column ‘winning_team’ was created with information from ‘score’ and ‘opponent_score’ in the original data frame by taking both values and showing the end result instead, as well as indicating how many points the winning team won by. Additionally, we needed to create a Twitter developer account to input these tweet IDs and be able to receive information about the tweet in response. The preprocessing of this data revealed that much of the data, being six years old, had become unavailable either because the tweet had since been deleted or the user chose to make their tweets private. As a result, much of the beginning of this project was focused on updating our collection of data to reflect what we could access as well as how much we need. Due to the significant size of these datasets, running through all of it would take significant amounts of time on our laptops so we decided to narrow down our focus to two teams: one very successful and one incredibly unsuccessful in terms of Super Bowl winnings. Next, we were able to narrow down which tweet IDs returned a tweet successfully and only keep those. By collecting the text of the tweets, we then updated the data frame with two new columns, one for the text of the tweet and one for the tokenized text to enable chart and graph making. By simply taking counts of word frequencies, it was obvious that further preprocessing was needed. We compiled a list of punctuation, some specific to tweeting (“…”, “#”, etc.), as well as a list of stop words adjusted to include project-specific words such as “RT” standing for retweet.

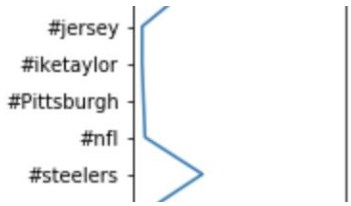
The beginning of our approach includes importing the csv files as six separate data frames, which are then separated by team, pulling out only the rows we need to make the run time as efficient as possible. At the beginning, we used only two teams, so we started with four data frames for ‘browns postgame’, ‘browns pregame’, ‘steelers postgame’, and ‘steelers pregame’. The original CSV team column was organized using two or three letter abbreviations for each team, which was relatively straightforward to figure out once we had the full team name and city, but it was not consistent in using just the city or the name of the team. Using a try-except loop, the team data frames are run through to find the tweet IDs that return a valid response and those are appended to the list as their ID while the IDs that return an error in the except part of the loop are added as “N/A”. This is to ensure the rows stay with their proper IDs when a new data frame is made of the rows with valid IDs only. This new data frame is made by using an if-else statement within a for loop that runs through the ‘valid_tweetIDs’ column, adding every value that does not equal “N/A”. The problem that occurred at this step was the limit the Twitter API imposes where a developer can only send a certain number of requests in a 15 minute period; as a result, we were required to wait 15 minutes every time we ran this step before being able to move on and run it again to create another data frame. However, in our first experiments with this step, it would take about forty minutes to run through more data and end up with less information, so with this new for loop format, it ended up being less cumbersome in comparison. For cleaning up the tweet text, we removed stop words, a list we got from NLTK, along with a few added words including “RT” as well as unnecessary punctuation also compiled in list form. To separate out the content words, we ran a nested for loop with an if-else statement to run through the column, each tweet, and then take only the words that do not occur in the lists

by using ‘if not w in’ the list. This new version of each tweet was added as a new column named ‘contentwords_only’. Then, to simplify the original “score” and “opponent_score” columns, an if-elif-else statement is used within a for loop to run through the current data frame values and determine who won the game or if it was a tie. These results along with a difference value are appended to two separate lists, “winning_team” and “score_difference” to be added to the new data frame, created by taking certain rows from the original team data frame. Next, the polarity score is added to the data frame by using TextBlob to return a value from -1 to 1 based on the tokenized words column ‘contentwords_only’. This is done by running the tokenized column of the data frame through a for loop and adding results from the TextBlob method “tb(tweet).sentiment” to a list of numbers representing polarity and subjectivity. Then, using another if-elif-else statement, this list of polarity and subjectivity is run, only using the polarity value, to convert these numbers into “positive”, “negative”, and “neutral”, each decision being appended to a list that becomes a column called “polarity_yesno” in the data frame. This process is done for each team before and after games until there are four data frames including information about tweet ID, team and opponent, week, the winning team, the score difference, the actual tweet text, the polarity score, and the simplified polarity yes or no answer.

After formatting the data frames in an accessible way, we analyzed the data in a couple exploratory ways using simple count graphs and looking at highest and lowest counts, but also using logistic regression and graphing polarity with week numbers/wins and losses. In our exploratory analyses, we saw many words used at a high frequency but no trend in exactly which words were often used except an obvious trend in words relating to the team itself, as shown by tweets about the Browns in Figure 1 and the Steelers in Figure 2.

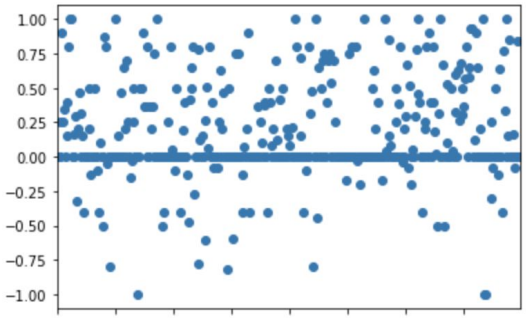


(Fig. 1 Browns pregame)

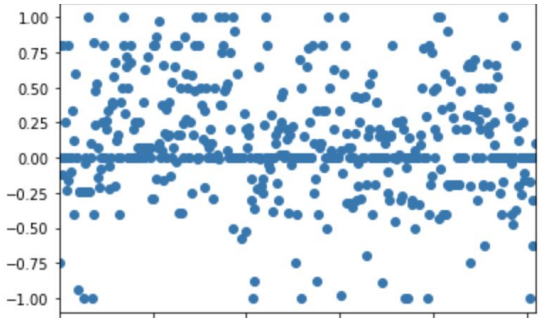


(Fig. 2 Steelers pregame)

Although there was a lack of trend in this surface level analysis, once we added polarity rankings the graphs became much more comprehensible. There is no definable change in polarity throughout the season, but the comparison of total negative and positive polarity reinforces our hypotheses that different success levels correlate with a different opinion on Twitter.



(Fig. 3 Steelers)



(Fig. 4 Browns)

As shown in Figures 3 and 4, the Steelers' success is reflected in an overall more positive leaning in polarity than the Browns receive on Twitter, and as mentioned before, the Browns have shown less success in their entire history, 2010 not being an exception. However, the Steelers do receive more neutral comments in general, making the proportion of positive to negative seem significantly higher than the Browns while it is not as large a difference in reality. Still, the phenomenon shown in the two figures below reflects a difference in Twitter's response to teams of different calibers. We were able to create a logistic regression model for predicting polarity depending on a win or a loss based on data from the postgame data frame and a linear regression model for predicting a win or loss depending on polarity based on data from the pregame data frame.

Pseudo R-squ.:	0.009442
Log-Likelihood:	-615.60
LL-Null:	-621.47
LLR p-value:	0.002829

(Fig. 5 Steelers Logit)

With the low LL-Null value from the Steelers' logistic regression model (Figure 5), we can reject the null hypothesis that polarity is not affected by the results of the game. To assess how accurate our hypothesis is, we looked at the Pseudo R-squared value which is .009, which shows that our model is not particularly strong but has merit. The linear regression model did not yield any viable results, no values for the Jarque-Bera test, f-statistic, Log-likelihood, and coefficients were significant enough to prove a correlational relationship between the polarity of Twitter before a game and the results of that game.

Pseudo R-squ.:	0.04526
Log-Likelihood:	-229.81
LL-Null:	-240.70
LLR p-value:	3.043e-06

y=winning_team[home win]	**coef**
Intercept	0.1207
polarity_score	1.3665

(Fig. 6 Browns Logit)

In figure 6, some stats from our logistic regression model are shown. The log values contribute to our decision to reject the null hypothesis that the result of a game does not have a correlation with the polarity of tweets about the game afterwards. The coefficient value being .1207 and closer to 0 than 1 makes a home loss the more likely result, another reflection of the Browns tendency to have fewer wins than the Steelers. In conclusion, our original hypothesis that game results do affect polarity of tweets in the following week was accepted while the hypothesis that tweets before a game could affect the game's outcome was rejected by the results of our regression models and polarity graphs.

Works Cited

Data collected from: Noah's Ark at http://www.cs.cmu.edu/~ark/football/