PREDICTING NFL OUTCOMES USING TWEETS

Kobe Mandell and Charles Stanier

Motivate the Problem

We wanted to design a net capable of predicting NFL game outcomes using fan tweets sent in the week leading up to the game as well as some baseline team statistics. Our goal is to see if the fans of teams have impacts or insights that affect the outcomes of games and to show that a network using Twitter data can make more accurate predictions than one using statistics alone, proving that there was some insight into the game to be found on social media.

Many other forms of predictive research acknowledge that 'the wisdom of crowds' or the aggregate of many predictions, each without their own individual accuracy guarantees, can combine to make fairly accurate predictions. Here, our crowd is the community of NFL fans on Twitter and their wisdom is their game-related tweets. This is interesting because in the social media age, 'wisdom' or at least people's opinions and predictions are widely found all across the internet. If this wisdom can be utilized to make effective predictions about relatively unimportant events such as sports games, it stands to reason that they could also predict more consequential outcomes like political elections or stock market fluctuations. And since the average person's online engagement is only going to increase in the years to come, the applications for systems driven by aggregated opinions are only growing.

Describe prior work in this area and explain why it does not address the problem

We have found research where a logistic regressor is trained on Twitter data and NFL statistics to predict outcomes of games [1]. These researchers found some success in predicting NFL outcomes using Twitter data and a combination of Twitter and NFL data.

We also found research where an artificial neural network was trained on only NFL statistics to predict NFL outcomes [4]. This group also found some success in their methods, which we will discuss later on.

We decided to take the main ideas from these two papers to answer our initial questions. How well can deep learning predict NFL outcomes, and can a machine learning framework achieve higher performance in predicting NFL outcomes when given Twitter data.

Describe your approach in very high level terms (What kind of learner did you use?)

We used a simple linear feedforward artificial neural network with two hidden layers as described in [4]. The neural network takes in a feature vector and outputs a vector of size 2. The first number is the probability the home team loses, and the second number is the probability the home team wins. The numbers add up to one. We used the Adam optimizer and binary cross entropy loss, as the two classifications were win or loss depending on which probability was higher.

We created two networks described above. The only difference is one network had 11 features in the input vector that was composed of NFL statistics and the second network had 15 features in the input vector. These features were 11 of the same NFL stats as before plus 4 extra Twitter features that describe the tweets about the matchup.

Describe how you tested and trained it

We gathered 3 years of NFL data to obtain traditional NFL statistics. Each example represented one NFL Game. The first 5 values of the feature vector represented home scoring, home passing yards, home rushing yards, home fumbles and home interceptions. The second 5 values were the same but for the away team. The last input feature was a home field advantage parameter.

We processed the NFL stats as described in [4]. For each NFL statistic, we found the average value up until that week. We also kept track of how many points, passing yards, rushing yards, fumbles and interceptions each team gave up. Thus, if the home team averages 10 points a game, and the away team typically gives up 20 points a game, the first value of

the feature vector (home scoring) would be 15 points, which is the expected home team scoring.

Using the Python package sportsreference, we obtained all of the data described above from 2010 - 2012. We omitted games from week 1 and week 17. We couldn't predict games for week 1 because it's the first week so there is no prior data on the teams. We did not use pre-season data as it is not representative of regular season competition. Similarly, we omitted week 17 because it is the last week of the NFL season and the competitiveness of those games is not representative for various reasons.

In the NFL there was only one tie over the years 2010-2012 and we omitted the example from our dataset. There are 32 teams in the NFL and 15 total weeks that we used. But that only means 14 games per team because each team has a bye week. That equates to 224 total games per year, so 672 total games over 3 years minus the one tie. Therefore our total data set contains 671 NFL Games.

We found a Twitter data set from [1]. The authors of the paper created their own dataset and posted it here, for academic purposes. The dataset was created by obtaining tweets from Twitter with specific hashtags. The researchers wrote out a list of specific hashtags for each team and if a tweet contained any one of those hashtags, the tweet was added to the data set. Depending on the time of the tweet, the opponent would be deciphered as the next opponent for the team that was tweeted about. In total the dataset has on average 500,000 tweets per year.

We took each Twitter ID from the data set and ran a script to get the content of the actual tweet. We fed the tweet in the Vader sentiment analyzer, which outputs a number to measure the polarity of the tweet on a scale from -1 to 1, where -1 is a bad sentiment and 1 is a good sentiment. We then averaged all of the tweet's polarity for a team with their respective matchup. We also calculated the percent change in volume of tweets from week to week, essentially tracking how much following a particular game got. We took that percent change and multiplied that by 0.2 as our "rate" feature. We used 0.2 because according to the research done in [1] it seems that 0.2 is the empirical optimal value for getting the best results. We added these 4 total features (2 for home and 2 for away) to the input vector.

In both nets we used a batch size of 29 for the training set and batch size of 7 for the validation set. We also made our nets deterministic for replication purposes. We adjusted the max_steps parameter to where the loss and accuracy of the nets started to level off and to avoid overfitting and overtraining. In both nets we split up the 671 NFL Games into 435 training examples, 104 validation examples, and 132 test examples, using a random but deterministic split of the data.

Sources of Error/Noise in the data

In the pre-populated list of hashtags from [1], not every tweet that contains one of those hashtags is about the actual team. In other words, a team might have posted about a Chicago hotdog and added a hashtag #gochicago. This tweet would be in our dataset because of the hashtag even though the tweet has nothing to do about football. Thus, our average polarity scores and tweet rates for each team are slightly noisy because some tweets are completely irrelevant and do not represent aggregate sentiment of a football matchup.

Results

For the stats net, we ended up running the net for about 9000 steps. Training Loss and Validation decrease throughout and Training Accuracy is rather stagnant while Validation Loss increases.

We found that our Twitter net learns much faster than the stats net alone. We only ran this net on only 1000 steps and we found loss and accuracy to both level off after a certain amount of steps. It is clear that the Twitter net was able to learn much faster than the net with only stats.

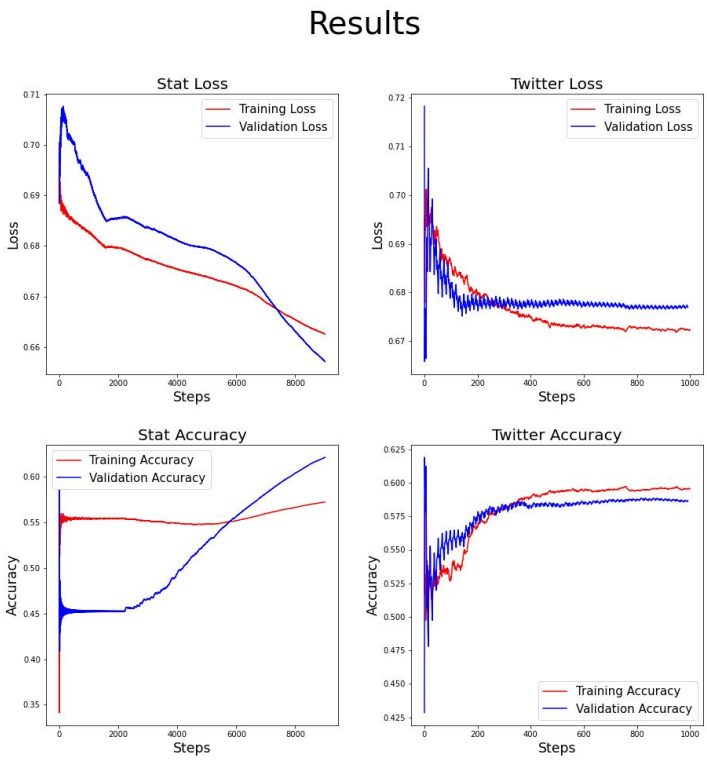
Metric	Steps	Final Validation Accuracy	Final Test Accuracy
Stats Net	9000	83.3%	60%
Twitter Net	1000	66.6%	65.9%

In the stats net the final validation accuracy was higher but the final test accuracy was lower. The stats were not able to generalize as well as the Twitter net which had consistent results between validation and test accuracy. This is potentially due to overtraining the net.

Thus, it is clear that after running on a much smaller amount of steps the Twitter net was able to have more accurate and consistent results.

Obtaining 65.9% over the testing set is slightly better than the performances of comparable classifiers. The research described in [1] had accuracy around 63.8% accuracy over their testing set.

The specific loss and accuracy curves from our experiments are shown below.



Future Work

There are a couple of concepts that we would like to research in our future work. One idea is to gather sentiment analysis of all forms of social media. Instead of just using Twitter, we would be able to explore how Instagram, Facebook, Snapchat, and Reddit give us a better measure of aggregate sentiment. Furthermore, we would be able to explore which social media platform gives the most representative sentiment of the population, and perhaps different social media platforms are better at predicting different events.

Moreover, our group would like to explore using different sentiment analysis packages and potentially combining sentiment analysis scores when predicting aggregate sentiment. An outlier sentiment score that has an error would be averaged out, and it would not make an impact on our dataset.

Lastly, we would also like to explore predicting stock price fluctuations based on aggregated social media sentiment and posts on LinkedIn.

References / Related Papers

[1] Shiladitya Sinha , Chris Dyer , Kevin Gimpel , and Noah A. Smith, "Predicting the NFL Using Twitter (2013)", *Carnegie Mellon University, Toyota Technological Institute at Chicago,* http://www.cs.cmu.edu/~nasmith/papers/sinha+dyer+gimpel+smith.mlsa13.pdf

[2] Blaikie, David, Abud, Pasteur, "NFL & NCAA Football Predictions using Artificial Neural Networks", *College of Wooster,* http://personal.denison.edu/~lalla/MCURCSM2011/4.pdf

[3] Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha, "Comparing and Combining Sentiment Analysis Methods (2013)", *Proceedings of the first ACM conference on Online social networks* https://dl.acm.org/doi/pdf/10.1145/2512938.2512951?casa_token=i42LT6lZ-PkAAAAA:5AJNEfN5qLC-9wlTRwIgPVgmMXDBotBmlQGwu8PQCzcnLh5tecReURIaj5BHzrc1_XVzx95smyWf

[4] John A. David, R. Drew Pasteur, M. Saif Ahmad, Michael C. Janning, "NFL Prediction using committees of Artificial Neural Networks", *Journal of Quantitative Analysis in Sports/College of Wooster,* https://www.degruyter.com/view/journals/jqas/7/2/article-jqas.2011.7.2.1327.xml.xml