Sentiment Analysis for Deflategate

Saurabh Kumar
Indiana University
Bloomington, IN 47405
kumarsau@iu.edu

Abstract

This paper implements a sentiment analysis model to process tweets regarding an American Football scandal, deflategate, and get the public sentiment for the whole ordeal. The model also does context based sentiment analysis to answer the specific questions regarding blame and support for each context. I have chosen three contexts, Patriots, Tom Brady and Management, to scrutinize the scandal. These contexts cover the entire breadth of the scandal as it incorporates the key points of the scandal. The model also calculates the overall criticism and support shown by people. The model classifies sentiment using Naïve Bayes Classifier.

1 Introduction

Deflategate is a NFL controversy regarding New England Patriots tampering with footballs used in the American Football Conference (AFC) Championship Game especially against the Indianapolis Colts on January 18, 2015. Before the 2015 Super Bowl, there was a good deal of coverage about a scandal around deflated footballs and whether the Patriots cheated.[1]

Tweets with #deflategate show people's reaction to the scandal. My model analyses these specific tweets to classify tweets into four categories, irrelevant, neutral, positive and negative. Irrelevant tweets show that people were not concerned by the scandal. Positive, negative and neutral tweets show how people judged the scandal.

The model preprocesses the tweet dataset to remove unwanted characters. Advanced processing is done on the data to make it suitable for the classifier. I have used a training tweet dataset[2] for the Naïve Bayes Classifier. The deflategate tweets become the train set and is fed to the classifier to predict its sentiment. Then model then searches for various contexts in the train data. Few comparison statistics are calculated on the contexts and sentiments. The various findings are shown by plotting various graphs for the comparison statistics.

2 Related Work

Although sentiment analysis is a growing area of study, not much sentiment analysis work is done in the field of sports. Sentiment analysis for any text classification does share some key points. Study in document level classification (Pang 2008) and learning polarity of words and phrases (Hatzivassiloglou 1997) has been done. The study of linguistics features to detect sentiments in tweet (Efthymios 2011) helps to for a base for sentiment analysis. Researchers have worked to find relevance of latest events and the sentiment from twitter data (Namrata 2007). Sentiment analysis for broader sentiment category has been done (Jon

[1] https://en.wikipedia.org/wiki/Deflategate

[2] Training Data: Sanders Twitter Sentiment Corpus. Link for the dataset: http://www.sananalytics.com/lab/twitter-sentiment/

2013). The broader categories include positive, negative, neutral and irrelevant tweets. In the recent past a lot of work is done in the classification of twitter data (Jansen 2009; Pak 2010; Barbosa 2010; O'Connor 2010).

In the area of sports studies include analyzing sentiments for sports tweet (Yang Yu 2015) finding sport sentiments (Alex 2007) and predicting tournament outcome using twitter (Shiladitya 2013).

3 Data

The main data is gathered from a public source[3]. Tweets regarding the scandal were recorded on the important days. I have reduced the dimension of the data to keep only the important fields. Data contains 9 attributes including the main information needed, tweets. The dataset has 11814 data rows.

This model uses a training dataset that contains 5513 manually classified tweets[4]. The tweets are classified into one of the four categories, irrelevant, neutral, positive and negative. This dataset has 5 attributes including *TweetTest* and their respective *Sentiments*. Since twitter restricts distribution of actual tweets with sentiment corpus, this training dataset is not openly available as a single file. I have used a small python script to download this data. The download time is about 43 hours because of the limitations in Twitter's API.

4 Method

The model that I have built can be described and divided into 5 major subparts. These subparts are shown in the architecture diagram given below. The square boxes show each part. The output of this model is a classified tweet data with sentiments and contexts. Sentiment classification is done using NaïveBayesClassifier() function in the nltk library of python. Context classification is done using automated keyword search in the tweets. Various graphs are plotted to better understand the results.

[3] https://www.crowdflower.com/data-for-everyone/

[4] Sanders Twitter Sentiment Corpus. Link for the dataset: http://www.sananalytics.com/lab/twitter-sentiment/

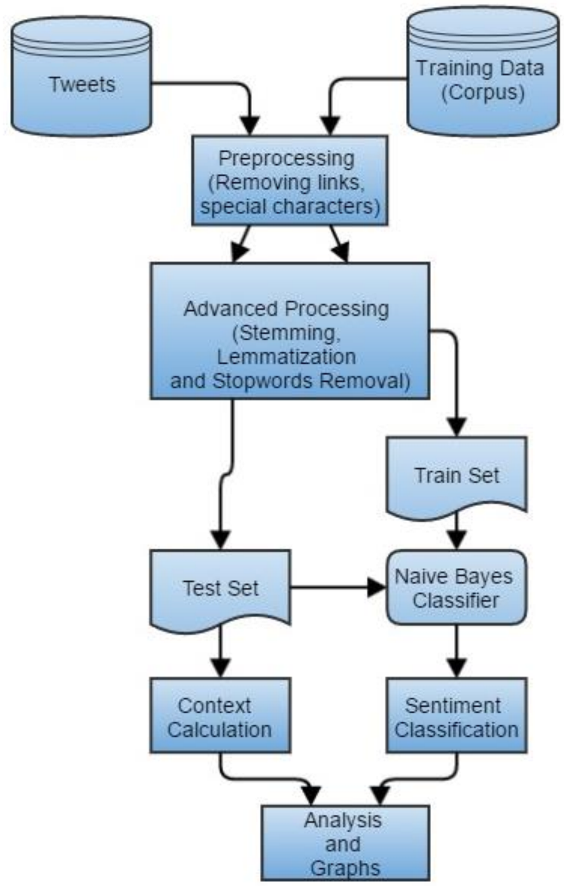

Figure 1: Architecture Diagram

4.1 Preprocessing

I have defined a preprocess() function to remove links special characters and non ascii characters from the tweets column in both the test and train dataset. This function uses regular expression library in python. It uses the sub() function to substitute the unwanted characters.

4.2 Advanced Processing

Three major tasks are performed in this stage, stemming, lemmatization and stopwords removal. Stemming means converting a word to its base form. Lemmatization is converting the word to its dictionary form. Stopwords[5] are the less important words that are removed before text processing.

Prior to performing these three tasks I have tokenized the tweets in both the test and train data.

[5] https://en.wikipedia.org/wiki/Stop_words

This helps in further processing and is done using word_tokenize() function in the nltk libarary of python. Stemming is done using PorterStemmer() function in nltk library. There is a stopwords dataset defined in nltk.copus. This dataset is used to remove stopwords. Lemmatization is done using WordNetLemmatizer() function in nltk library.

4.3 Sentiment Classification

For sentiment classification, I have created a classify_naivebayes() function which takes training and test dataset as inputs. The training set contains tweets and their respective sentiments. As discussed earlier sentiments are classified into 4 categories. I have used NaiveBayesClassifier() function from the nltk library. It learns through the training set. The test data only contains tweets. This data is fed to the classifier to predict the classification of each tweet in the test set. The final output is the tweets and their respective sentiments for the test data.

4.4 Context Calculation

Contest calculation is done by the find_context() function which takes test data as input. Context is calculated for each tweet by a simple search process. I have built this function to search for specific contexts, Management, Tom Brady and Patriots. The tweets that don't have the specific keywords are classified as others. The output of this function is the context relating to a specific tweet.

4.5 Analysis and Graphs

For this stage I have created two functions, analysis() and graph(). This method takes input the sentiment and context list created by the previous two functions. It calculates 5 statistics, which are dictionaries in python.

Sentiment_stat is a dictionary with 4 key value pair with the 4 types of sentiments as its key and the number of sentiments as its value. Blame_stat does the same for 3 different types of contexts, with only the tweets with negative sentiments taken into consideration.Support_stat uses the positive tweets to do the same. Tom_stat contains 2 values, the number of negative tweets and the number of positive tweets, with Tom Brady as the context. Pat_stat does the same with patriots as the context. These 5 statistics are passed to the graph() function to plot respective graphs. I have plotted histograms for each statistics using the matplotlib library in python.

I have created a word cloud for the tweets in the test dataset to show the important words in the deflategate tweets. This is done using WordCloud() function in the wordcloud library of python.

5 Evaluation

The result of my model is shown through various graphs for the 5 statistics. These graph help to answer few key questions regarding the scandal.

Figure 2 shows the sentiment statistics. We can see from it that the majority of tweets were negative tweets. Neutral and positive tweets are almost equal in number. Irrelevant tweets are low in number. This graph shows that most people had negative review for the scandal.

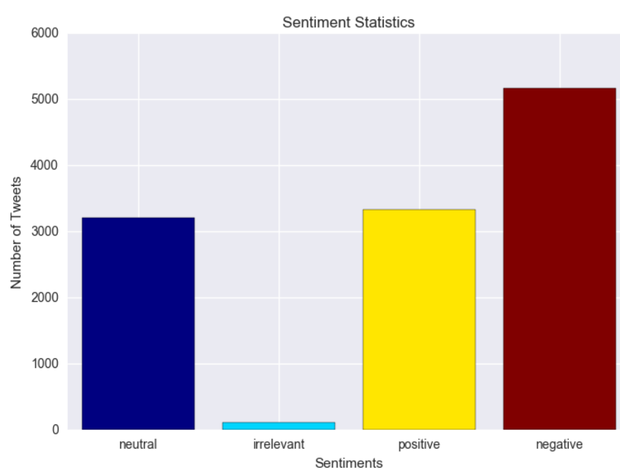

Figure 2: Sentiment Statistics

The figure below shows the blame statistics for the scandal. The player, Tom Brady was blamed the most followed by the team, Patriots. Few people blamed management for this scandal.

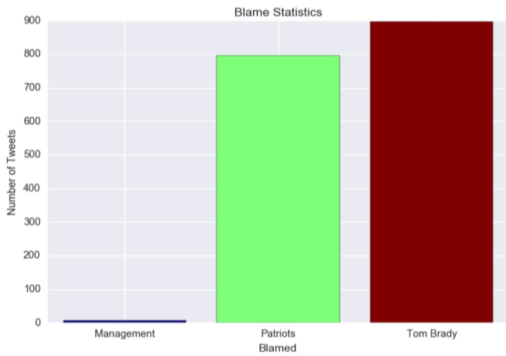

Figure 3: Blame Statistics

Figure 4 shows the support statistics for the scandal. Most of the people with positive review about the scandal supported Patriots. Support for Tom Brady is a second close.

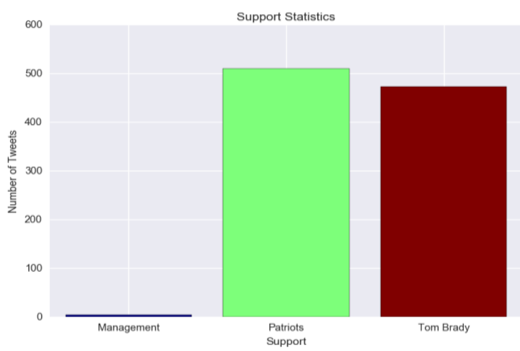

Figure 4: Support Statistics

The figure given below shows that more people blamed Tom Brady. The blame number is twice as much as the support number.

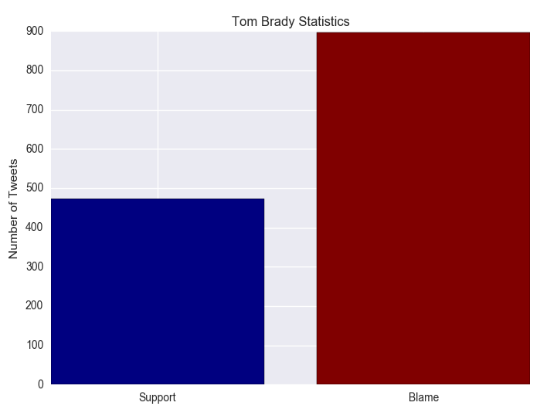

Figure 5: Tom Brady Statistics

Figure 6 shows that Patriots got more blame than support. This is consistent with the fact that most people had negative review about the scandal. The difference in blame and support is not as much as we saw for Tom Brady.

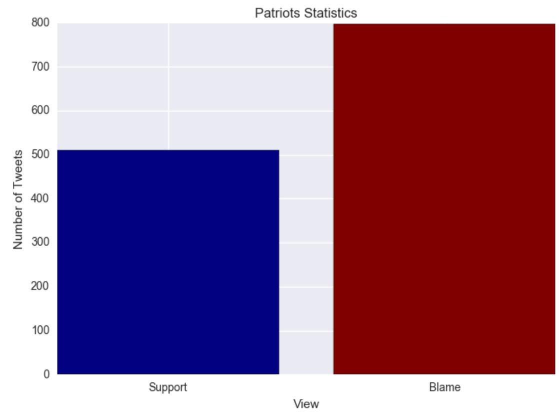

Figure 6: Patriots Statistics

The word cloud given below shows the most used words in the tweets. The font of the word in the figure is correlated to the frequency of that word. So words like Deflategate, Patriots, Brady and NFL are frequently used in the tweets.

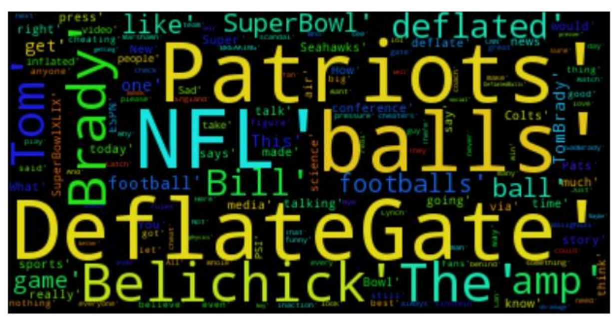

Figure 7: Word Cloud

6 Future Work

Though the model built by me completes its goals, there is still a lot of room for optimization and improvement. Firstly, I can improve the training set by adding data rows related to this specific scandal. This will require manual classification. It helps to better train the classifier. Additional contexts could be added to answer more questions. The context calculation part can be done using a classifier like Naïve Bayes. Alternate classifiers, like SVM and SLDA, can be used for sentiment analysis. A comparison study can be done to show the best classifier for this task and eventually the model can incorporate the best classifier.

7 Conclusion

The model successfully provides a gist and analysis of the deflategate scandal from the data present in tweets. It answers important questions regarding the scandal. As we have seen earlier people mostly have negative review regarding the scandal and they mostly blame player Tom Brady for it. The team, Patriots, was also heavily criticized. In spite of all the blame, people showed some support towards the team, Patriots. A large chunk of people were neutral about the scandal.

References

Efthymios Kouloumpis; Theresa Wilson; Johanna Moore 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG!, Proceedings of the Fifth International AAAI Conference.

Namrata Godbole; Manjunath Srinivasaiah; Steven Skiena 2007. Large-Scale Sentiment Analysis for News and Blogs, ICWSM.

Jon Tatum; John Travis Sanchez 2013. Twitter Sentiment Analysis. Alternation. Stanford University, US.

Hatzivassiloglou, V., and McKeown, K. 1997. Predicting the semantic orientation of adjectives. In Proc. of ACL.

Jansen, B. J.; Zhang, M.; Sobel, K.; and Chowdury, A. 2009. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology 60(11):2169–2188.

Pak, A., and Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In Proc. of LREC

O'Connor, B.; Balasubramanyan, R.; Routledge, B.; and Smith, N. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of ICWSM

Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In Proc. of Coling.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2):1–135

Yang Yu 2015. World Cup 2014 in the Twitter World : A big data analysis of sentiments in U.S. sports fans' tweets, Computers in Human Behavior.

Alex Edmans; Diego Garcia; Øyvind Norli 2007. Sports Sentiment and Stock Returns, The Journal of Finanace Vol LXII, No. 4.

Shiladitya Sinha; Chris Dyer; Kevin Gimpel; Noah A. Smith 2013. Predicting the NFL Using Twitter. CMU, US