Data Mining Final Project

1. An **abstract** that summarizes your work - highlight the QUESTION you're trying to answer

2. An **introduction** that motivates the problem you are trying to solve.

3. Section(s) describing your **methodology** - what did you actually do

**4. Results** and/or **evaluation** section(s), with data or figures to support your claims as appropriate.

5. A brief **future work** section explaining what is left to do.

6. Appropriate citations and **references** from the literature.

Abstract:

Question: Are tweets a better estimator for election outcome than polls, and if so, what is the best method for utilizing tweets in order to predict election outcomes?

Are tweets a better estimator for election outcome than polls, and if so, what is the best method for leveraging tweets to predict election outcomes?

Introduction:

Methodology:

*Data Collection:*

In order to compare Twitter data to the polling data, I first had to collect the data. I ran a script that collected all tweets the contained the string “Election2016” between July 10, 2016 and November 7, 2016. I collected a total of 856,985 tweets, which don’t include retweets or replies. I then filtered these tweets to only look at the ones containing information about one candidate at a time. After this filter, I had 268,249 tweets total. A comparison of tweet counts to polling data counts can be seen in **Figure 1.**

I found the raw polling data online in csv format. This data contained 304,496 total responses between the dates of July 10, 2016 and November 7, 2016. Before comparing this data to the Twitter data, I first had to adjust the values to represent a two party race. I used the formula

APC1 = PC1 / PC1 + PC2

where PC1: percentage for candidate 1, PC2: percentage for candidate 2 and APC1 is the adjusted percentage for candidate 1 [2]. The adjusted values from the final outcome of the election can be seen in **Table 1**. After the data were uniform, I was able to graph them together to compare the differences, which can be seen in **Figure 2**.

*Sentiment Analysis:*

With each tweet containing only one ‘topic,’ I could then directly apply the sentiment to that tweet to the candidate it was referring to. The sentiment analyzer I used was called Vader Sentiment, which “is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains [1].”

This specific sentiment analyzer returned a ‘normalized, weighted composite’ score that was computed “by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). This is the most useful metric if you want a single uni-dimensional measure of sentiment for a given sentence [1].”

I counted a tweet as a ‘vote’ for the corresponding candidate if it had a sentiment score >= 0.05, and I counted the tweet as a vote for the other candidate if the sentiment score was <= -0.05. Scores in-between were counted as neutral, so neither candidate received a vote.

*SVM Model:*

I then trained two models in order to determine if a model trained on the polling data or Twitter data performed better. I trained two SGD Classification models with L2 regularization and alpha values of 1e-4 on the polling data, as well as the twitter data. The polling classification model was trained on 80% of the polling data and the twitter classification model was trained on 75% of the data.

For the Twitter data, the tweets were first converted into a Bag of Words model using scikit-learn’s CountVectorizer method [3], then it was passed through the TfidfTransformer which, “transforms a count matrix to a normalized tf or tf-idf representation. Tf means term-frequency while tf-idf means term-frequency times inverse document-frequency [3].” The testing and training error from these models can be seen in **Table 2**. The results from predicting the final election outcome using the tweets from November 7, 2016, can be seen in **Table 3**.

*Analysis of Bias:*

The following equation was used in order to calculate bias,

BC1 = APC1s – APC1f

where BC1 is the bias towards candidate C1, and APC1f is the actual election outcome (see **Table 1** for actual election outcomes) for C1 and APC1s is the adjusted percentage of C1 in the poll currently being analyzed [2]. The bias was only calculated against one candidate (Trump in this case) since BC1 = -BC2. All biases shown are against Trump.

The calculated bias of the predicted outcomes from the SGD model are shown in **Table 3**. I then calculated the bias of the polls from each day, and compared that to the bias from the tweets from each day. This comparison can be seen in **Figure 3**. The average poll bias was 2.1603 whereas the average Twitter bias was 1.2064. The polling data had a range of biases between [-1.6085, 5.2758] which is a range of 6.8842. The twitter data had a range of biases between [-26.1426, 20.9761] which is a range of 47.1187. I then tried calculating the election outcomes from the tweets by counting each sentiment, regardless of day, to each candidate as a ‘vote’ and then found calculated the bias of those estimations, which are shown in **Table 4.**

Results:

Insert Graphs Here

Conclusion:

Future Work:

I was only able to find the raw data for one polling site, but it would be more ideal to obtain the data from many different polls, and average them together. I would like to also analyze the sentiments from the Twitter data from further before the election, and also after the election to see if the general sentiment has varied. I would also like to analyze specific tweets where the difference in sentiment between the candidates is large, and try and pinpoint the exact event(s) that took place at that time to create such drastic differences in sentiment. I would also like to add more features/instances to the polling dataset to see if that would improve the model.

References:

[1] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014**.**

[2] Anuta, David, et al. “Election Bias: Comparing Polls and Twitter in the 2016 U.S. Election.”*Election Bias: Comparing Polls and Twitter in the 2016 U.S. Election*.

[3] [Scikit-learn: Machine Learning in Python](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.