Predicting Elections through Twitter

Akhil Jain
2011CS50274

Sharad Maheshwari
2011CS50295

Abstract

This study explores the usage of sentiment and time series information expressed in a tweet about candidates in the US Presidential Primaries of 2016. It is observed that simple methods are able to predict election results to a reasonable level of accuracy. This points to the opportunity for leveraging Twitter as a metric of user sentiments towards electoral candidates and their correlation with the real results.

Introduction

Today, social media influences the thoughts and actions of the people to a large extent; so much so that people post online asking for advice and opinions on a wide range of issues. At the same time people use these platforms to express their opinion on everything that happens around them. Twitter is a popular micro blogging website with over 300 million active users that enables users to send and receive short 140 character messages called tweets. Tumasjan et al. (2010) have already shown that people use microblogging to exchange information about political issues. In this light, we explore the extent of insight we can gain into the political views of an individual on Twitter and predict the results of a forthcoming election based on tweet analysis.

Background on the US Presidential Primaries

The party candidate of the two major political parties in the US - the Republicans and the Democrats for the Presidential elections in the US is selected through a series of presidential primary caucuses and elections held over a period of 5-6 months (Wikipedia 2016). Thus, there is a series of 50+ elections between the same group of contenders within a short period of time.

In the 2016 primaries, there are two main contenders for the Democrat seat - Bernie Sanders and Hillary Clinton; while for the Republican seat, there are four main contenders - Donald Trump, John Kasich, Marco Rubio and Ted Cruz.

Related work and research questions

Since 2010, there has been much work on predicting the vote shares that the candidates get in an election. Dwi Prasetyo and Hauff (2015) have shown a comprehensive survey of the literature in this field. The common metric used to evaluate the effectiveness of a model in these papers is the Mean Absolute Error (MAE) which is defined as:

$$MAE = \frac{1}{n}\sum_{i=0}^{n}|predicted_i - actual_i|$$

where n is the number of candidates in the election.

A large number of shortcomings of these approaches have been pointed out in Gayo-Avello (2012), fundamental amongst which is the criticism that most of this work is done on a *post hoc* basis, that is, after the election has already been completed. Therefore, most of these models suffer the fundamental flaw of overfitting to the results of a particular election. This is most evident from the fact that the original paper of Tumasjan et al. reports an MAE of 1.7% in the German Elections, while the same approach when applied to the Nigerian Presidential elections, has an MAE of 11% (Fink et al. 2013) and when applied to our dataset of US primaries has an MAE of 18.5%.

Some work has also been done on using sentiments for more accurate prediction of tweeter's loyalties in the election (Bermingham and Smeaton 2011). However, the problem with these approaches is that most of them require hand labelling and therefore are time consuming and not scalable. Also, the results of these papers are not significantly different from the results obtained from the non-sentiment based approaches.

Therefore, the goal of the present study is to propose a generalised and automated method for the prediction of the winner in an election, with special focus on the US Presidential Primaries, 2016.

Dataset collection

In order to predict elections using Twitter, we will need to collect tweets over the representative period which correspond to the political event of concern. However, we forsee two major issues with this approach:

1. If we look at tweets from around the world on US politics, it will generate a lot of noise since almost everyone in the world has something to say about and is influenced by happenings in the US political space.

2. If we restrict ourselves to the tweets containing the geotags of the place where the election has taken place on a particular date, we run into constraints such as too few tweets because most people tweet without sharing their locations.

In order to obtain the right balance between the two, for each candidate, for two weeks preceding the date of the primary, we collect all tweets with geotags within the state where the primary will be held and 650 tweets from across the globe per day. From the global corpus, we remove the tweets which refer to more than one candidates because they would unnecessarily lead to inclusion of noise in the data. Thus, we collect (by scraping the web pages using Selenium) a total of 4439 location specific tweets and 303130 global tweets.

Methodology

We use the following 2 baselines for this problem:

1. MAXSOFAR: In this approach we simply predict the candidate having highest wins so far as the winner for election being considered.
2. COUNTS: We use the idea promulgated in Tumasjan et al. (2010), that is, set tweet share (fraction of tweets containing the name of the candidate) as indicative of vote share and then declare the one with maximum share as winner of the election. In this way all the elections and their predictions were independent of each other.

Instead of just using the fraction of tweets that a candidate garners as representative of the vote share, we believe that the sentiment expressed in the tweet about the person mentioned in the tweet is also extremely important. For example, a tweet that says *"Hillary Clinton sux"* clearly suggests that the tweeter has a negative view about Hillary Clinton and is most probably not going to vote for her but will contribute to her tweet share above. Thus instead of tweet share, better features would be tuples of $(positive, neutral, negative)$ shares for the tweets.

We also believe that the time series information will be critical in determining the results. This is so because the candidate with greater positive opinion closer to the election should perform better than the candidate with greater positive opinion 2 weeks out. In order to test these hypotheses, we try the following two models:

1. BAGSENTIMENTS: In this, we obtain 6 features per candidate per election - 3 features from the sentiment tuples generated from the location specific tweets and 3 from the tweets that have been collected globally.
2. TIMESENTIMENTS: In this, instead of using all the global tweets as a single bag, we generate 3 features per day from the sentiment tuples and concatenate those over the 14 days preceding the election. Thus we have a total of 42 global features and 3 local features (there aren't enough local tweets for us to be able to split further on the basis of days) per candidate per election.

We then concatenate all the feature values for each candidate in an election and train standard multi-classification learners [1] to predict the winners. It should be noted that as on May 01, 2016, a total of 41 Republican and 40 Democrat Primaries had been completed and therefore the number of training examples is quite small. Thus, in order to offset the disadvantage and prevent over fitting, we used 10-fold validation and trained multiple models with parameter tuning to obtain the optimal model. Eventually, it was found that the Gaussian SVM gives the best performance.

Sentiment Analysis

In order to create a target-independent target-based tweet sentiment analyser, we use ideas presented in Jiang et al. (2011) in addition to some general tweet sentiment analysis features such as tweet normalisation, removal of URLs and user mentions, expansion of clitics etc. We have a human annotated dataset of 2538 tweets for Bernie Sanders, Donald Trump and Hillary Clinton (annotated by students as a part of a class project) and another dataset of 1.6 million general tweets annotated by the presence of emoticons which we use for training. Each occurrence of a candidate's name is replaced by a dummy variable X to neutralise the tweet's bias towards the candidates and then the associated features are used to train sentiment classifiers. We then train three models for sentiments - one for the general tweets, one for the person specific tweets and one which combines the predictions of the aforementioned models into a single prediction.

It should be noted that sentiment analysis in itself is a hard problem in Natural Language Processing and is way off from being solved because of which we might encounter some errors. However, these methods obtain f-scores of around 0.5 on political tweets.

Results

Since we do not want our predictions to be biased, we do all the predictions using 10 fold validation i.e. train on 90% of the total data and then test on the remaining. We then compute accuracy and macro F score for all the baselines and our improved models using these predictions.

Republicans

For the republicans, as can be seen in Figure 1 TIMESENTIMENTS performs best with a Macro F score of 0.41 and an accuracy of 63.41%.
Other measures for the republican data using our approach are as follows:

- Macro Precision: 0.58
- Macro Recall: 0.31

Democrats

For the democrats as well, we see in Figure 2 that TIMESENTIMENTS achieves the best performance with a Macro F score of 0.56 and an accuracy of 55.00%.
Other measures for the democrat data using our approach are as follows:

[1] http://scikit-learn.org/stable/

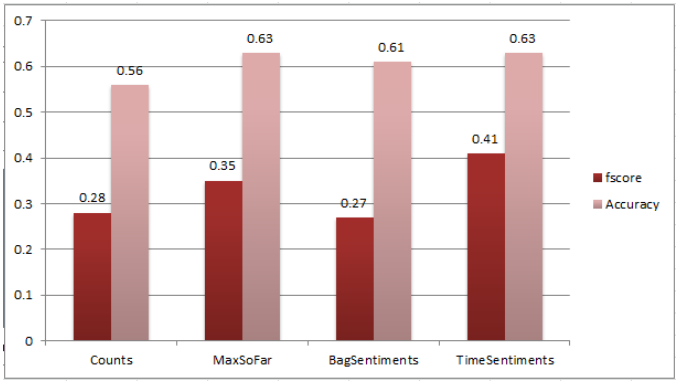

Figure 1: Republicans

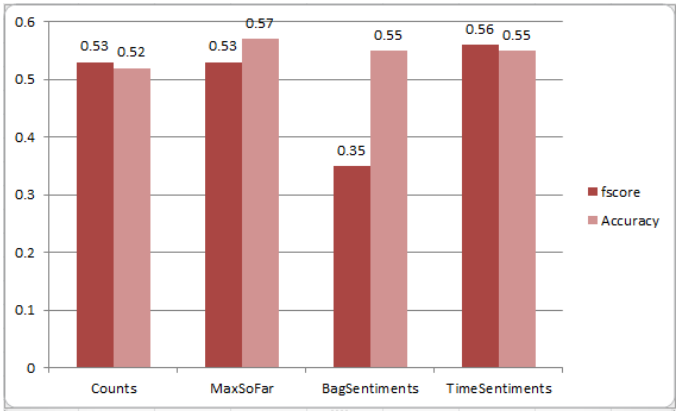

Figure 2: Democrats

- Macro Precision: 0.56
- Macro Recall: 0.56

Conclusions

Thus from the results above, it can be seen that Twitter sentiment is a better predictor of elections than mere count of mentions of the election candidates. Also, it is seen that time series information provides better insights into peoples' political viewpoints.

Despite the results, we would like to point out that this study suffers from multiple limitations, such as:

1. The electorate's complete demographics may not be represented on Twitter. Though the level of technology penetration in a country like the US is expected to be high, there would still be more young people than old on Twitter and therefore an accurate representation of the demographics would not be obtained. A possible way to counteract this effect in the future is to incorporate the demographic information into the feature space just like we have done for the time series information.
2. Target based sentiment analysis still has a long way to go before we can make accurate predictions about the sentiments expressed about a target in a tweet.
3. The dataset we have collected could contain multiple tweets by the same person, therefore giving greater weight to one person's viewpoint than others. Similarly, we would have missed out on tweets that are in reply to tweets about a candidate but do not explicitly mention his name and all of the ensuing discussion. We would need ways of including cascade information in the feature space to be able to handle these.

In summary, despite the many limitations, it is seen that the TIMESENTIMENTS model performs reasonably well on the dataset and is able to predict the results of the election better than all the other models that we have tried out. The fact that a fairly simple method tried out over a short period of time is able to generate plausible results on the dataset bears testament to the power of predictions using Twitter data and thus possibilities to leverage the technology.

References

Bermingham, A., and Smeaton, A. F. 2011. On using twitter to monitor political sentiment and predict election results.

Dwi Prasetyo, N., and Hauff, C. 2015. Twitter-based election prediction in the developing world. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, 149–158. ACM.

Fink, C.; Bos, N.; Perrone, A.; Liu, E.; and Kopecky, J. 2013. Twitter, public opinion, and the 2011 nigerian presidential election. In *Social Computing (SocialCom), 2013 International Conference on*, 311–320. IEEE.

Gayo-Avello, D. 2012. " i wanted to predict elections with twitter and all i got was this lousy paper"–a balanced survey on election prediction using twitter data. *arXiv preprint arXiv:1204.6441*.

Jiang, L.; Yu, M.; Zhou, M.; Liu, X.; and Zhao, T. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 151–160. Association for Computational Linguistics.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM* 10:178–185.

Wikipedia. 2016. United states presidential primary — wikipedia, the free encyclopedia. [Online; accessed 4-May-2016].