Comparing other methods of predicting election outcomes

Prediction has always been the talk of the town when popular events are going to take place. In the latest clash between boxing legends Floyd Mayweather and Manny Pacquiao, they even had boxing legends of previous decades to come together to provide expert analysis on which fighter would win the boxing match. In the end Floyd Mayweather won. It would not have surprised anyone who believed in numbers because Floyd had a perfect winning streak of 47-0 while Manny had 5 losses in his fight career before the fight. We have chosen to predict the election results for the US because it is the next big thing.

**Our approach**

We will be crawling a corpus out of Reddit and Twitter for the comments made about two presidential candidates; Donal Trump and Hillary Clinton. This will be done by our python scripts which will call upon available APIs from Reddit and Twitter.

Reddit is an online community which allows its users to post any comments and vote on content. [1] According to the company’s page, it reported on 15 September 2015 that in August 2015, it had 202,818,688 unique visitors from 208 different countries. [2] In addition, US president Barack Obama uses it to reach out to his supporters. [1] As such, we believe that Reddit is a good resource to predict the winning outcomes of US presidential candidates.

Twitter is a social media which allows users to post or reply via tweets. Tweets are text messages and have a limit of 140 characters. Tweets can be viewed by anyone regardless of being a member or not. [3] Twitter currently has 316 million of monthly active users and 23% of the accounts reside in the US. If we take 23% of 316 million users, we have 72.68 million users which is still a substantial amount. [4]

After gathering the corpus, we would use sentiment analysis to separate positive comments from negative ones. Instead of just analysing text, we will include emoticons. We will then compare the ratios of positive and negative comments between the two candidates. Our hypothesis is that the candidate with a higher ratio of positive comments will stand a higher chance of winning the election.

**Other approaches**

Phone polling

Phone polling was conducted by polling companies in Britain. They collected a sample of 1,000 people arbitrarily and weighted the samples according to demographics such as age, location and gender to represent the population. Polling companies apply different weights on the samples based on certain criteria. An example of criteria would be asking the participants’ past voting choice and use this information to weight their voting intentions accordingly. The biggest limitation using this approach is that is it unsustainable in the future even though it proves to be more accurate than online polling most of the time. Internet polling is less costly and can gather more information from more people. As such, it would be a better representative of the population. [5] Hence, in the near future, phone polling will most likely be replaced with online polling.

Data science

Nate Silver is one of the famous statisticians who shot to fame in 2008 when he forecasted the outcome of the primaries and presidential victors in 49 states. [6] He is the founder of FiveThrityEight website which publishes political articles and his predictions. [7] The success of his prediction was attributed to many factors. Nate Silver gathered data from a very broad scope ranging from demographics to economic variables. The crucial factor was his ability to choose the correct regression models with complex statistical modelling software. [8] It may seem that Nate Silver’s method of prediction is the answer for all political predictions. However, his predictions were far from accurate for the UK election this year. In addition, his predictions became less accurate since 2012 US presidential election. [9] As such, there is a need to find an alternative solution to predict the political results.

Research papers

Similar research has been done to analyze comments from twitter to predict the outcome of elections. In this research, the way it does its sentiment analysis is by assigning a sentence to be positive if there are any positive words in the sentence. As such, in the case of a sentence with a positive word and a negative word, it can be both a positive and negative statement at the same time. The lexicon used is from OpinionFinder. From the sentiment analysis, the two presidential candidates which it compared were Obama and Mccain. Since the campaign can only have one winner, it is expected that the sentiment for one candidate would vary inversely from each other. However, they seem to slightly correlate in the sentiment analysis. [11]

In this research paper, it found other sources that proved traditional social media to be a reliable option to predict election outcomes. However, the same thing cannot be said for twitter because tweets are a mere 140 characters. In addition to this, a market consultancy even said that 40% of the tweets are “pointless babble”. This research paper focused on German election. The methods used were downloading tweets in German and then translating it to English. Thereafter, it used the LIWC2007 (Linguistic Inquiry and Word Count) to assess the emotional components of tweets. The research found out that twitter though was dominated by a small number of heavy users, the tweet volumes is close to the results of federal election. [12]

This research paper acknowledges the usefulness of using tweets to predict certain things like movie successes but not so much for elections. It uses the algorithms from [11 & 12] on the 2010 US Senate special election in Massachusetts to prove that the success of predicting the elections is a coincidence because it was not repeatable on another data set. There are a few possible reasons for this failure to predict. One of them is the manipulation by spammers. Fake accounts can be easily created and by spamming positive remarks on a certain politician can distort the view of any observer. [13]

This research paper agrees with [13] and took a step further by including more information from the tweets such as geo location of each user. It has two algorithms. The first one gets the location of the user through the location field while the second one checks the confidence of the predicted location based on the contents it received from algorithm one. Algorithm 2 is necessary because users sometimes key in irrelevant information such as ‘bedroom’ as the location. At the end of it, it concluded that it is feasible to predict American presidential elections using tweets but there are several limitations. One of them is the current programs does not integrate the dynamics of political conversations in social media. [10]

Another research paper suggested that a web-derived lexicon will bring about a tremendous improvement on a lexicon-based sentiment classifier. [14]

So it seems that the current technology or level of research fails to predict the outcome of elections consistently. As such, we do not plan to come up with a new algorithm which ambitiously aims to predict accurately our all data sets due to the time constraint. Hence, our team has decided to investigate on the contents of social media which were not included in these researches. We will be focusing on the effect of emoticons to better classify (positive or negative) the comments made by users.

[1] Christy Loerzel, “What is Reddit and why should you care?” [Online] Available: <http://www.symantec.com/connect/blogs/what-reddit-and-why-should-you-care> [Accessed 29 Oct 2015]

[2] Reddit, “Reddit About”, [Online] Available: <https://www.reddit.com/about/> [Accessed 29 Oct 2015]

[3] Tech Target, “Twitter” [Online] Available: <http://whatis.techtarget.com/definition/Twitter> [Accessed 29 Oct 2015]

[4] Twitter, “Company Facts”, [Online] Available: <https://about.twitter.com/company> [Accessed 29 Oct 2015]

[5] Alberto Nardelli and Tom Clark , “Meet the pollsters who are predicting the general election results” [Online] Available: <http://www.theguardian.com/politics/2015/apr/08/meet-the-pollsters-predicting-general-election-results> [Accessed 29 Oct 2015]

[6] LeighBureau, “Biography” [Online] Available: <http://www.leighbureau.com/speakers/NSilver/> [Accessed 29 Oct 2015]

[7] Nate Silver, [Online] Available: <http://fivethirtyeight.com/> [Accessed 29 Oct 2015]

[8] David Smith, “How Nate Silver won the election with Data Science” [Online] Available: <http://blog.revolutionanalytics.com/2012/11/in-the-2012-election-data-science-was-the-winner.html> [Accessed 29 Oct 2015]

[9] Dylan Byers, “Nate Silver: Polls are failing us” [Online] Available: [<http://www.politico.com/blogs/media/2015/05/nate-silver-polls-are-failing-us-206799> [Accessed on 29 Oct 2015]

[10] Lei Shi, Neeraj Agarwal, Ankur Agrawal, Rahul Garg, Jacob Spoelstra, “Predicting US Primary Elections with Twitter” [Online] Available: <http://snap.stanford.edu/social2012/papers/shi.pdf> [Accessed 07 Nov 2015]

[11] Brendan O’Connor, Ramnath Balasubramanyan, Bryan R. Routledge, Noah A. Smith, “From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series” [Online] Available: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1536/1842/> [Accessed on 29 Oct 2015]

[12] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe, “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment” [Online] Available: <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/download/1441/1852> [Accessed 07 Nov 2015]

[13] Daniel Gayo-Avello, Panagiotis T. Metaxa, Eni Mustafaraj “Limits of Electoral Predictions Using Twitter” [Online] Available: <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2862/3254> [Accessed 07 Nov 2015]

[14] Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, “The viability of web-derived polarity lexicons” [Online] Available: <https://www.aclweb.org/anthology/N/N10/N10-1119.pdf> [Accessed on 29 Oct]