CASE STUDY: ELECTION ANALYTICS ON TWITTER

We will do a simple case study on using Twitter to gain key insights into important social and political events such as elections. In particular, we use the example of the US 2020 presidential election in this case study.

Opening prompt

Twitter, one of the most influential online social media, has 145 million daily active users and 500 million tweets posted per day. Millions of tweets are posted on Twitter, expressing users' emotions, opinions, and promotion of cultural and political views.

Hashtags, a metadata tag that is prefaced by the # symbol, play a crucial role in microblogging and photo-sharing services such as Twitter. People widely use hashtags to express what they want to convey. It can be an event or a name of a public figure, such as a famous singer or even a politician. From hashtags, researchers can mine lots of information. Twitter can be used to understand trends, such as who is poised to win a presidential (e.g., U.S. 2020) election. How do we combine text analysis and hashtags to do election analytics and what kinds of analytics should we be aiming for?

Context and basic steps

The first step is to obtain the tweets from the Twitter streaming API. One way to do so is by using Tweepy, which allows programmatic access to tweets from Twitter. The search words used to crawl the data are five hashtags: #DonaldTrump, #JoeBiden, #2020Election, #Vote, #Debates2020. We collected around 274 thousand raw data from October 21st to November

8th. After crawling the data, we got the JSON structure data, which is shown in Figure 1. It contains User IDs, DateTimes, hashtags, and URLs.

One kind of simple analytics we can perform is *hashtag prediction.* Since many tweets have hashtags, we can use these for training, and the rest for prediction. This allows us to separate tweets that seem to be talking about Biden (for example) from those talking about Trump. Next, we would want to understand whether Trump, Biden or any other candidate (such as the third party candidate, Jill Stein) are being referred to positively or negatively. We can try to approximate this by using sentiment analysis. One good example is VADER, which is the Valence Aware Dictionary and sEntiment Reasoner. VADER not only tells about the Positivity and Negativity score butalso tells us about how positive or negative a sentiment is. Vader also gives us a compound score. The compound score is a metric that calculates the sum of all the lexicon ratings, which have been normalized between -1 and 1, which is from most extreme negative to most extreme positive. We define the compound score larger or equal to 0.05 as positive sentiment, less or equal to -0.05 as negative sentiment, and the compound scores between are defined as neutral.

Once we have the sentiment analysis scores, as well as hashtag predictions, we can try to get a sense of how support for Biden, Trump or Stein are fluctuating over time.

Case Questions

i) What are some of the cons of using twitter for election analytics? Hint: consider the prevalence of bots, as well as user bias.

ii) What are key differences between data preprocessing on twitter and 'normal' text such as news articles? What extra steps would you take?

iii) Other than hashtag prediction, what other analytics can you think of where you can get training data from a subset of Twitter, but then use the trained model to analyze unlabeled data from the rest of Twitter?

iv) Would Twitter be a good fit for doing similar analytics for the US midterm elections? Why or why not? Hint: Is every political figure running for election known or popular in mainstream social media? What about the ones who have heavy presence on social media themselves?

v) How do you validate what you're seeing from Twitter analytics? Could you evaluate sentiment analysis results using polling data, for example? What obstacles do you see? Hint: Tweets may or may not have location metadata enabled. What problems can this cause for elections? Would similar problems arise if you were instead trying to evaluate something more politically neutral such as a marketing campaign for a product or service?

vi) What are some ways in which you can visualize the results of hashtag prediction and sentiment analysis? What are some metrics for success?