Group 10: COVID-19 Tweets by World Leaders

Intermediate Report

Anh Tran
Gray Buckley
Jakob Horvath
Tegan Tingley

KEYWORDS

datasets, natural language processing, text tagging

ACM Reference Format:
Anh Tran, Gray Buckley, Jakob Horvath, and Tegan Tingley. 2018. Group 10: COVID-19 Tweets by World Leaders: Intermediate Report. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 2 pages.

1 PROJECT CHOICE

For the project proposal, we proposed using Twitter as our data source. However, we had three project ideas that we proposed. We have decided to pursue our first project idea - analyzing COVID-19 tweets by world leaders and then comparing the language used in these tweets.

2 TWITTER HANDLES

In order to begin this project, we needed to define a list of world leaders and find their Twitter handles. Due to issues with translating other languages into English and maintaining the meaning of the Tweet, we decided to limit our queries to Twitter accounts that tweeted in English at least most of the time. We made a list of countries around the world and compiled Twitter handles for high-ranking leaders in those countries. For example, for the United States, we decided to query tweets written by:

- President Biden
- Former President Trump
- Nancy Pelosi, Speaker of the House
- Mitch McConnell, Former Senate Majority Leader
- Andrew Cuomo, Governor of New York
- Ron DeSantis, Governor of Florida

While Andrew Cuomo and Ron DeSantis do not hold positions in the federal government, they received a lot of media attention during the pandemic. For other countries, we included official government Twitter handles, such as @10DowningStreet the official UK Prime Minister account. After this process was complete we had a list of XX Twitter handles to query.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00

3 OBTAINING DATA

We initially planned to use Python's TweetPy library, however due to limitations with this library, we instead decided to use the Python library *snscrape*. This can be used to scrape information from various different social media platforms, including Twitter, Reddit, Facebook, and Instagram. For our purposes, we will only use it to scrape data from Twitter. [1]

4 FILTERS

While COVID-19 was a major news topic of 2020, this was not the only major news story of the year. Hence, we needed to filter tweets to tweets related to COVID-19. We created a list of words related to the pandemic and then used an OR statement in our query. Here are some of the COVID-19 related words used in our filter.

- COVID
- Mask
- Vaccine
- Coronavirus
- Viral
- WHO
- CDC
- Fauci

5 QUERY

Using *snscrape*, we scraped data from Twitter. As mentioned in the previous section, we used a filter to query tweets based on COVID-19 related words. We pulled the date/time of the tweet, tweet ID, user, tweet text, number of likes, number of replies, number of retweets, and the country. We pulled the number of likes, replies, and retweets in order to be able to assess the impact of the tweet.

A snapshot of some of the data pulled is shown below in ADD FIGURE.

6 WORD TO VEC

After querying all of the tweets by world leaders about COVID-19, our next step was to get the data in a usable form for clustering. We decided to use Word to Vec to help handle this natural language processing task. [2]

7 NEXT STEPS

We plan to perform clustering on the Twitter data. We plan to use K-Means++, however, we may need to try to use a couple different clustering algorithms if we run into issues.

The purpose of clustering the data is to try to group words together and then determine which world leaders used those words

and how frequently they used the words. This may help us identify relationships between countries, governments, and leadership.

We would also like to use the number of likes, replies, and retweets to potentially analyze the impact of tweets by world users. We may need to create some sort of baseline by country or leader - i.e. what is a common number of retweets - in order to effectively study this. Tweets with higher than average replies, likes, and retweets will likely have a higher impact than tweets that are below average.

REFERENCES

[1] https://github.com/JustAnotherArchivist/snscrape
[2] https://www.tensorflow.org/tutorials/text/word2vec