An analysis of 20.000 mini stories

Data Wrangling – Project Report

By: Berend Markhorst, bmt800

Research Question

What kind of language and tone does Donald Trump use in his tweets and how does this relate with the world close around him?

The main question actually consists of 2 questions. The first question, concerning language and tone, can be answered using the following sub questions:

1. Which words/combinations of words does Trump use most frequently?
2. Does the location, time or the device Trump uses affect the tone of his tweets?

The second part of the main question, concerning the world close around him, can be answered using the following sub questions.

3. What kind of relation is there between Trump's tweets and google trends?
4. What kind of relation is there between Trump's tweets and stock prices of DJI or CYN?
5. What kind of relation is there between Trump's tweets and the weather?
6. How does an activity as playing golf influence Trump's tone and language in his tweets?

Data Sources

Data Source	URL	Python module	Last Access Data
Twitter API	https://twitter.com/realDonaldTrump	Tweepy	18-01-2019
Google Trends	https://trends.google.nl/trends/?geo=NL	pytrends	18-01-2019
Yahoo Finance	https://finance.yahoo.com/	fix_yahoo_finance	18-01-2019
Historical Weather Data Washington DC	https://www.meteoblue.com/en/weather/archive/export/washington-d.c._united-states-of-america_4140963	Read via pd.read_csv()	18-01-2019
Golf Data	https://trumpgolfcount.com/displayoutings	Read via pd.read_csv()	18-01-2019
Github: Twitter Data 2014-2018	https://github.com/bpb27/trump_tweet_data_archive	Read via pd.read_csv()	18-01-2019

Data Wrangling Methods

- Every tweet comes as a JSON-file from Twitter. From all recent tweets, only the relevant information was taken and put in a dataframe. Then, stop words, punctuation and typos were removed from the tweets using the python module NLTK. The remaining part of the tweets were

stored in a list in a separate column. Finally, also nicknames/synonyms were changed to the original form of a word (e.g. “dems” becomes democrats) by hand.

- An important aspect of a tweet is the date/time at which it is tweeted. In the JSON-file, they were stored as strings. That is not useful; they had to be changed to datetime objects. To make plotting easier, the date and time were split into “date only” , “time only” and “part of the day”. The same had to be done for Trump’s golf data.
- The tweets’ sentiments were calculated with the python module VaderSentiment.
- All aforementioned methods were also applied on the Twitter data archive from 2014 to 2018.
- To compute the most used combination of words, all tweets were divided into tuples with neighbouring words. This was done using “ngrams” yielding digrams.
- For both Google Trends, Yahoo Finance and the historical weather data from Washington DC, dataframes were merged with columns of Trump’s dataframe. Sometimes, NaNs were dropped. Sometimes, NaNs were replaced by 0 (e.g. when Trump does not tweet about a certain subject on a certain day). In case of stock prices, sometimes stock prices from the day before (e.g. for holidays) were assigned to a certain day.
- For plotting, matplotlib was used. Mainly box-, bar, pie, scatter and regular plots were used to visualize data. Sometimes seaborn was used. Also WordCloud was used for visualizing the amount a word is used in Trump’s tweets.

Conclusion

1. In the last 2 weeks (first 2 weeks from January 2019), Trump mentioned “border” and “border security” the most. In the last 4 years he mentioned “Trump” the most, followed by “great” and “president”. In the top 10 of words he uses a lot, there’s only 1 positive. The rest is neutral.
2. For security reasons, the location has been muted on Twitter. The president uses an iPhone and has used an Android phone. His Android-tweets tend to be more positive. During the day, he also becomes more and more negative.
3. There seems to be a relation between, for example, the number of times someone looks up “border” on Google and the number of times Trump mentions “border” in his tweets.
4. There does not seem to be a relation between the stock price of the DJI/CYN and Trump’s tweets.
5. Heavy rain seems to stimulate the president to talk nicely on Twitter. However, short, little rainstorms seem to result in a drop in the president’s sentiment.
6. Playing golf does not result in more positive tweets. Even more, it looks like the president is more agitated on Twitter after playing golf.

In general, the distribution of Trump’s Twitter sentiment tends to be left skewed. Therefore, he seems to be relatively positive in his tweets.

Limitations

Cynicism, something that president Trump masters and often uses in his tweets, does not get noticed by the sentiment analysis. This is because that algorithm only focuses on the words rather than on the context. This could be improved in future research.

Social Media are not a good indicator of someone’s emotions. Often, the same person differs quite a lot from his/her online version. The same holds for the president. Looking at press conferences from him in the last 2 weeks, he does not seem positive at all while his tweets say so. For future research, one could ask whether Twitter is a reliable source concerning someone’s emotions.