A CANDIDATE'S TWEET DURING THE PANDEMIC

How did Joe Biden engage with Twitter population during the Covid-19 outbreak?

Elena Zordan
M.Sc. Data Science, University of Trento
Computational Models of Human Behavior
Instructor: Bruno Lepri

ABSTRACT

This study aims at analysing the tweets of the actual president of USA Joe Biden during elections months. The purpose is to see how the candidate engaged with the citizens via Twitter during the coronavirus outbreak through sentiment analysis. The analysis is performed through natural language processing tools such as NLTK and NRClex.
The results show that Biden adopted an overall winning approach, reaching an increasing number of supporters throughout months.

KEY WORDS

Sentiment analysis, Biden, NLTK, NLP, text mining, covid-19, Twitter, sentiment-based dictionary

INTRODUCTION

2020 was characterised by several exceptional events that revolutionised the world and reshaped people's life in ways never seen before. Among others, also US presidential elections were conducted in very different ways from past editions, bringing to light new questions and needs, which leaded candidates to modify their approaches to engage with citizens.

Even though a shift in this context already began back in 2008, when Barack Obama started to use social media as a medium that "allowed incumbents and newcomers alike to speak directly to constituents on everything from policy to what they had for dinner" (Yildirim, 2020), these last months forced

whoever wanted to be listened to opt for this change.

In fact, when talking about important figures, the attention on these "informal" channels increases because of the potentially huge consequences and political implications. Nowadays, people are able to easily produce content and have their voices heard in ways that were inconceivable until two decades ago, and above all, systemic changes within the broader media ecology like the expansion of choice in the media environment, lead to an increasingly important role of social networking sites as sources of political information (Theocharis and Jungherr, 2020). The conjunction of these elements is continuously transforming traditional political information production, distribution, and consumption dynamics.

That is why this project aims at studying social media data of Joe Biden's twitter account throughout 2020. The purposes of this work go beyond the scope of an explanatory data analysis of Biden's tweets: the goal is to understand the approach by which the candidate chose to engage with his potential voters, analysing tweets' content through opinion mining techniques, especially by means of sentiment analysis methods.

On the other hand, the personal reason behind the choice to focuses on the democratic candidate rather than the republican opponent Donald Trump, was dictated by the huge amounts of studies about this last one. In addition to that, in the time of Covid-19, nations all over the world face not just a major public health crisis, but also a crisis of social relations. Especially in settings of entrenched inequalities and political polarization, the pandemic has exposed and exacerbated conflicts between social groups (Uyheng and Carley, 2020). Finally, Twitter was preferred to other social media since it has emerged as a major digital public relations tool for politics over the last decade, besides being well equipped with APIs and NLP libraries.

The paper is structured as follows: after the introduction, the literature section concerns sentiment analysis methods in a computational social science context; data and methodology paragraph lists the techniques used in the

analysis, after that, results and a final discussion conclude the work.

LITERATURE

Computational social science is defined as the "development and application of computational methods to complex, typically large-scale, human behavioral data" (Lazer *et al*, 2009).

The discipline encompasses a wide range of areas and techniques, it aims at using large archives of naturalistically created behavioral data (e.g., emails, tweets, Facebook contacts) to answer to social relevant questions.

Among the other methods, a relevant position is covered by those under the name of text mining.

Text mining is defined as a technique capable of exploring large amounts of unstructured text data in order to extract information and patterns.

Therefore, it allows to make better informed decisions, automate information-intensive processes, gather business critical insights and mitigate operational risks. Indeed, the are several applications of text mining, such as sentiment analysis, speech recognition, spam filtering and e-commerce personalization. Among others, sentiment analysis is now considered as the most iconic one, actually it "has become an interesting field of study in its own right [and this is also due to the] availability of digital data about opinion statements" (Veltri, p. 178, 2020) that can be gathered from multiples sources such as product reviews, social media platforms, blogs and so on.

It is described as a "technique to gauge the sentimental content of a writing [that] can help understand attitudes in a text related to a particular subject" (Rajput *et al*, p.2, 2020). Belonging to the natural language processing (NLP) techniques and classifying emotions in textual data, it allows to analyse and predict the polarity of sentiments within a text, either positive or negative. Therefore, sentiment analysis is often performed on subjective data such as emails, survey responses, social media data and beyond.

Sentiment analysis was first introduced by Liu in 2010 in his study entitled *Sentiment analysis*

and subjectivity. Handbook of natural language processing (Yunis, 2015).

With respect to related works, many researchers have worked on sentiment analysis on Twitter social media in literature, in particular with respect to Covid-19 tweets and Donald Trump tweets.

Kaur and Sharma, in *Twitter Sentiment Analysis on Coronavirus using Textblob* (2020), studied the sentiments regarding coronavirus by analysing the sentiments of different people's opinion for this disease. Tweets were collected through twitter APIs, then they were related to positive, negative and neutral emotion by the use of machine learning approaches and tools. In addition, for pre-processing of fetched tweets NLTK library and Textblob were used (Manguri *et al*, 2020).

M. Ra, B. Ab, and S. Kc, in *COVID-19 Outbreak: Tweet based Analysis and Visualization towards the Influence of Coronavirus in the World* (2020), visualized the influence of Covid-19 by executing algorithms and methods of machine learning in sentiment analysis on the tweet dataset in order to understand very positive and really negative opinions of the ultimate public round the world. The study revealed that the LogitBoost, a blended approach, performed better than other classifiers, reaching an accuracy of 74% (Manguri *et al*, 2020).

On the other hand, Ouyang and Waterman, in *Trump, Twitter and American Democracy* (2020), focused on Donald Trump's tweets, where they found how the former president used social media very strategically. In particular, they noticed that the more negative the tone of his tweets, the more likely it was that Twitter users will retweet his messages, concluding the work by saying "the public thirsts for negative Trump tweets" (Ouyang and Waterman, 2020).

METHODOLOGY

The analysis is performed in Python on a Jupyter notebook through NLP tools such as NLTK and NRClex, while seaborn, word cloud and matplotlib are used for data visualization purposes. Natural Language Toolkit (NLTK) is a set of open source Python

modules used to work with human language data for applying statistical natural language processing.

NRClex is an affect dictionary capable of measuring emotional affect from a body of text, in particular it includes fear, anger, anticipation, trust, surprise, positive, negative, sadness, disgust and joy. NRClex contains almost 27,000 words, it is based on the National Research Council Canada (NRC) affect lexicon, and the NLTK library's WordNet synonym sets (Bird *et al*, 2009).

Data visualization is performed with specific python libraries such as matplotlib, word cloud and seaborn, which provide several kinds of visualizations and allows a quite satisfactory level of personalization.

Analysis flow

The process pipeline is divided into four main steps: exploratory data analysis (EDA), text pre-processing and feature extraction, sentiment analysis on dictionary-based methods, data visualization and interpretation.

Exploratory data analysis aims at showing some insights about data before any text pre-processing operation.

Text pre-processing consists of several procedures; first of all there is the tokenization, which is the process of "breaking up a stream of text […] into phrases, words, symbols or either meaningful elements called tokens, [hence its goal] is the exploration of the words in a sentence" (Veltri, p. 175, 2020).

Then, text pre-processing involves stop words removal, where words that have little meaning such as "a", "and", and so on are removed, and rare words removal.

After that, the token normalization reduces words to base form to be analysed as single term; in order to do this, stemming and lemmatization are performed. While names are different, these two operations are very similar since they both remove and replace suffixes to get to the root form of the word; however lemmatization uses vocabulary list and morphological analysis to get the root word, hence lemma is an actual language word whereas a stem might not be one.

Text pre-processing part is concluded by part-of-speech tagging (POS) and Named Entity Recognition (NER). The former marks words in the corpus to a corresponding part of speech tag based on its context and definition; the latter seeks to extract a real-world entity from the text and sorts into predefined categories such as names of people, organizations, locations and so on.

Supervised sentiment analysis section is based on the use of vocabulary in which words are rated in terms of emotional "scores", in fact dictionary-based methods are all centred around the determination of a text's average happiness (sometimes referred to as valence) with sentiment dictionary through an equation, as described by Reagan and colleagues in his work (Reagan *et al*, 2017).

In order to better assess the approach adopted by Joe Biden towards the pandemic via Twitter, dataset is then split into three time periods according to the evolution of the outbreak: December 2019 - March 2020, April 2020 - August 2020 and September 2020 - November 2020.

Data

The dataset chosen is a csv containing Joe Biden tweets from 2012 to November 2020.

The csv contains also the number of likes, quotes, replies and retweets for each tweet.

The dataset is available at kaggle.

For the purposes of this project, the relevant tweets are those created from the end of 2019 onward. The reason behind this choice is due to the fact that coronavirus became a daily topic since then, when the world naively thought it to be mainly a Chinese problem.

Therefore, a subset of the original csv was created, which takes into account only tweets subsequent to December 2019.

RESULTS

Results were quite consistent with the expectations: the consensus toward the democratic candidate increased significantly during 2020, and this can be seen also from the initial plots which show almost an exponential growth of number of likes. Moreover, while at a first glance coronavirus does not seem to be one of the main topics of the candidate (when

counting the most frequent words), a deeper analysis revealed that the number of words related to Covid-19 outbreak is actually very high, reaching the third most frequent "word". The words that stand out more than anything else are always "Donald", "Trump", "Donald Trump", which highlighted that the majority of tweets were about Biden's republican rival. This should not surprise since the analysed tweets were gathered just before presidential elections.

However, from the twitter-logo word cloud other interesting words are noticed, such as "together", "health care", "family", "people", "work" and "support". In fact, these are consistent with the moral principles that characterize Biden's political campaign. In light of this, when analysing sentiment on tweets, the results confirmed an overall positive and neutral approach, rather than a negative one.

While the overall neutral and positive sentiment persists when the analysis is performed also in different periods of time, the same does not happen with respect to the nine NRClex emotions. In fact, strictly positive sentiments are consistently more that the strictly negative ones; but, while disgust is always the least present emotion, others change their position, such as joy, which increased significantly over months.

Further analysis details are well commented throughout the notebook.

CONCLUSION

Overall, the study has answered to the underlying question of the project, which purpose was to analyse the tweets of the democratic candidate during the Covid-19 outbreak, in order to see how the subject chose to engage with citizens via Twitter. In fact, through the diverse set of available instruments able to analyse large amounts of textual data it is not only possible to "provide indicators that help scientists understand collective behaviour [and] inform public policy makers" (Reagan *et al*, p.1, 2017), but it is also quite easy to inform the public about policy makers behavior.

The work points out that Joe Biden opted for neutral and positive approaches, capable of inspiring hope and faith to citizens. Moreover,

even though the analysis reveals that greater attention was gave on elections rather than the Covid-19 outbreak, it is interesting to notice that the most retweeted Biden's tweet contains both a political and a coronavirus-related hints, besides showing a scientific-oriented position, which was not taken for granted at that time. Moreover, the number of Covid-19 cases in USA became urgent later than other countries. Also, it has to be said that sentiment analysis comes with its own limitations which hamper the progress and the accuracy of the models. In particular, sarcasm is the biggest challenge to cope with, besides the use of slangs, hashtags, and the specific context in which the sentence lies (Veltri, 2020).

The world of sentiment analysis and more in general of NLP is fascinating and full of potentialities, but further steps are needed. With respect to this study, it would be interesting to daily analyse tweets of the actual president of the USA and compare them with the former president, since the pandemic still seems to be an urgent and challenging problem for everyone.

REFERENCES

- Steven Bird, Edward Loper and Ewan Klein, Natural Language Processing with Python, O'Reilly Media Inc, 2009
- Kamaran H. Manguri, Pshko R. Mohammed Amin, Rebaz N. Ramadhan, Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks, Kurdistan Journal of Applied Research (KJAR), 2020
- David Lazer, Alex Pentland, Lada Adamic, Sinan Aral,Albert-László Barabási,
  Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, Marshall Van Alstyne, *Computational social science*, Science, 323: 721-723, 2009
- Yu Ouyang, Richard W. Waterman, *Trump, Twitter and the American*

Democracy, Palgrave Macmillan, 2020

- Yildirim Pinar, *How social media is shaping political campaigns*, Knowledge Wharton, 2020
- Nikhil Kumar Rajput, Bhavya Ahuja Grover, Vipin Kumar Rathi, *Word frequency and sentiment analysis of twitter messages during coronavirus pandemic*, 2020
- Andrew J Reagan, Christopher M Danforth, Brian Tivnan, Jake Ryland Williams, Peter Sheridan Dodds, *Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs*, EPJ Data Science, 6, 28:1-21, 2017
- Yannis Theocharis, Andreas Jungherr, *Computational Social Science and the Study of Political Communication*, Political Communication, 2020
- Joshua Uyheng and Kathleen M. Carley, *Bots and online hate during the COVID-19 pandemic: case studies in the United States and the Philippines*, Journal of Computational Social Science, 3:445–468, 2020
- Giuseppe Veltri, *Digital social research*, Polity Press, Cambridge, 2020
- Eman M.G. Younis, *Sentiment analysis and text mining for social media blogs using Open source tools: an empirical study*, International Journal of Computer Applications, 112, 5:44-48, 2015