Hanna Born, Thomas Malejko, and Nicole Yoder

Dr. Lisa Harper & Dr. Linda Moreau

ANLY 580 (Fall 2020)

27 October 2020

An Exploratory and Analytic Review of

2020 Democratic Presidential Nominees Tweets

***Motivation:***

Since its introduction in 2006, Twitter has evolved into a key platform for modern politics – offering politicians a fast and easy way to communicate their message and priorities to the public. Under the Trump administration, Twitter has continued to grow even more central to politics as the president has often chosen his Twitter account as the preferred medium for connecting with the American people. Twitter’s popularity as a medium for political discourse means that information from political candidates is readily available in real-time as they post. This project focuses on the tweets made by seven major Democratic Party Presidential Nominee hopefuls from the start of their campaign until Super Tuesday, addressing the following topics:

* General Tweeting Trends
* Tweet Topics
  + Topics of Conversation
  + Do the Topics Change over Time?
* Authorship
  + Hacked Account Detection
  + Who’s Tweeting? The Candidate or Staffer?

***Literature Review:***

The rapid expansion of social media over the past decade has challenged the efficacy of existing analytic techniques due to the enormous quantity of data generated by these services as well as changes to communication structures. Subsequently, much research has been conducted to understand the feasibility and techniques required to generate useable insights from this data. Four papers, in particular, look at the practicability of conducting topic modelling and authorship detection of tweets.

In the “Author Identification on Twitter,” Antonio Castro and Brian Lindauer show that an author can be detected with 40 percent accuracy using a regularized linear regression model that relies only on publicly available information on Twitter. Interestingly, the authors comment that tweeters can evade detection by deliberately altering “their writing voice, or limiting the amount of text posted,” which—while beneficial to the political dissenter that the author is concerned about—may complicate our attempt to identify the originating Twitter account since most politicians likely have a multitude of personnel writing their tweet—from public affairs staffs, to aides, to the candidate themselves. Castro and Lindauer’s work builds upon the preeminent research conducted by Arvind Narayanan, et al. in a paper entitled, “On the Feasibility of Internet-Scale Author Identification.” The latter authors showed that neural networks and regularized linear regression models perform equally well for authorship identification, once the data has been normalized, and developed enhanced evaluation metrics including improved confidence estimators. Brunna de Sousa Pereira Amorim, et al. researched classification techniques that could identify tweets that contained political content and, hence, potential electoral crimes in Brazil (for example, out-of-term political advertising is illegal). As with the previous studies mentioned above, they found that logistic regression models far outperformed neural networks in this task, correctly identifying political tweets with nearly a 90 percent accuracy. Finally, “An Evaluation of Topic Modelling Techniques for Twitter” evaluated multiple techniques for topic modelling on ‘short’ documents, of which Twitter is completely comprised. This paper showed that biterm topic models outperformed—as measured by coherence scores—Latent Dirichlet Allocation (including those modified for use on short texts) and word embedded models such as word2vec.

***Dataset:***

The dataset being used for this research was collected from 28 May to 9 June 2020 and is comprised of 13,814 tweets from seven of the most prominent 2020 Democratic Presidential Nominee hopefuls, including: Joe Biden, Pete Buttigieg, Tulsi Gabbard, Amy Klobuchar, Bernie Sanders, Tom Steyer, and Elizabeth Warren. The tweets span from 2 August 2019 (approximately the beginning of the 2020 Democratic Primary Campaign) to 2 March 2020 (Super Tuesday), a seven-month period wherein each candidate tweeted at least 1000 times. This dataset was collected via Twitter’s API, reformatted, and saved in a comma-separated format that is 5,646 KB in size. In addition to the full text of the tweet, the dataset also contains information about the time, retweet count, and number of times that each tweet was favorited, in addition to, information about the user’s account at the time of the tweet such as follower count and friend count.

***Plan (Objectives):***

Our project will be broken into three general sections:

* Exploratory Data Analysis (led by Hanna) will explore:
  + Similarities and differences between the candidates’ tweets
    - Average tweet length, word length, lexical diversity, vocabulary size
    - Vocabulary growth by tweet count
      * See if any candidates tend to be “broken records” that repeat the same words/phrases
    - Use of mentions and hashtags
    - Most common words for each candidate and the entire group
  + Similarities of the most popular tweets (most likes/retweets)
* Topic Modelling (led by Nicole) will explore:
  + What topics can be latently discovered for each candidate and the entire group?
    - May test various algorithms from the Jonsson and Stolee (2006) paper:
      * LDA, LDA-U, Biterm Topic Model (BTM), and a word2vec Gaussian mixture model
  + Was there a change in topics over time?
    - May split tweets into the first and second half (before and after 16 November 2019) and repeat topic modelling analysis
* Authorship Detection (led by Tom) will explore:
  + Does each candidate have a distinctive tweeting style?
    - Since their staffers are most likely writing tweets for them, is there still a common voice among the staffers and candidate?
  + Is it possible to detect if an account has been hacked?

This is our planned workflow:

* Coordinate with GitHub
* Complete initial work in separate .py scripts:
  + One script for cleaning/organizing data, which the following scripts will pull the data from:
    - One script for EDA
    - One script for Topic Modelling
    - One script for Authorship Detection
* Selectively pull sections from those scripts for the Jupyter notebook report

This is our planned timeline:

* Weekly meetings to touch base
* 24 November – complete technical requirements and start poster board
* 30 November – poster board review and rehearsal
* 1 December – presentation
* 7 December – report review
* 8 December – report due

***Expected Result:***

This project expects to gain insight regarding Twitter’s role in modern politics, particularly among the seven Democratic candidates for the time period leading up to Super Tuesday. The analysis should provide a better understanding of the commonalities and differences in tweeting habits among candidates, as well as a look at the breadth of online conversation generated by the candidates. Additionally, we will explore and assess the feasibility of using authorship identification to detect fraudulent activity or to predict whether a tweet was likely posted by the candidate themself or by a staff member/aide. Looking into these three areas should provide a more wholistic understanding of how each candidate relied on Twitter politically leading into Super Tuesday and about the role Twitter continues to play in modern politics.

***Works Cited:***

Castro, Antonio, and Brian Lindauer. *Author Identification on Twitter*. http://cs229.stanford.edu/proj2012/CastroLindauer-AuthorIdentificationOnTwitter.pdf. Accessed 22 Oct. 2020.

de Sousa Pereira Amorim, Brunna, et al. “Using Supervised Classification to Detect Political Tweets with Political Content.” *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web - WebMedia ’18*, ACM Press, 2018, pp. 245–52. *DOI.org (Crossref)*, doi:10.1145/3243082.3243113.

Jonsson, Elias, and Jake Stolee. *An Evaluation of Topic Modelling Techniques for Twitter*. https://www.cs.toronto.edu/~jstolee/projects/topic.pdf.

Narayanan , Arvind, et al. “On the Feasibility of Internet-Scale Author Identification.” *2012 IEEE Symposium on Security and Privacy*, https://people.eecs.berkeley.edu/~dawnsong/papers/2012%20On%20the%20Feasibility%20of%20Internet-Scale%20Author%20Identification.pdf. Accessed 22 Oct. 2020.