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: (HANNA DO)*

- Nominee Tweeting Tendencies\*\* Topics of Conversation

- Hacked Account Detection (Authorship)

- Trends of Mentions Who is Being Mentioned—people that should be apolitical?

*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):* (NICOLE LEAD)

-- Bulleted List

-- By Responsibility

-- Build Scripts and Selectively Pull to Jupyter

-- One Clean / Organize Script

-- One EDA – pull Clean Script

-- One Topic – pull from Clean Script

-- One Authorship -- pull from Clean Script

Exploratory Analysis (HANNA LEAD)

- Similarities & Differences Between Candidates Tweeting Styles

- Tweet Length, Word Length, Lexical Diversity, Vocabulary Size

- Inclusion of Mentions / Hashtags

- Common Words

- Similarities of Popular Tweets?

- Vocabulary Growth by Tweet Count? Broken Record Analysis

- Hashtag Analysis?

Timeline

8 December—Report Due

7 December – Report Review

1 December—Project Presentation

30 November – Poster Board Review / Rehearsal (RXL)

24 November – Technical Requirements Complete, Start “Poster Board”

Weekly Touch Points Otherwise

- Tweet Sentiment?

Tweet Topics That Generated Traction (Increased Follower Count?) (Increased Likes?)

🡪 Future Topic

Topic Modelling – Common to All Candidates? (NICOLE LEAD)

- First Half v. Second Half

Authorship Detection (TOM LEAD)

- Staffer or Actual Candidate Tags?

- Common voice among staffers?

*Expected Result: (HANNA DO)*

- Understand the commonalities and differences in tweeting habits by candidate

- Understand the breadth of online conversation generated by the candidates

- Assess the feasibility of using authorship identification to detect fraudulent activity

*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.