Politicians you can Trust:
Classifying Legislator COVID-19 Tweets for Reliability

Adam Bryant, Ashton DeRousse, Benjamin Klein, Timothy Krenz
University of Tennessee, Knoxville
Min Kao Bldg. 1520 Middle Dr
Knoxville, TN 37996

Abstract

The unprecedented COVID-19 pandemic has accelerated the need for the dissemination of accurate information on the novel virus to that of a public health concern. To this end, a set of COVID-19-related tweets from members of congress were classified by their reliability. This set was used to train a neural network to further classify over 100,000 COVID-19 related tweets from 505 members of congress to check for potential demographic predictors of reliability. Ultimately, tweets were somewhat successfully classified by NLP sentiment, and there are not significant demographic predictors of a member's reliability. Classifying reliability on sentiment alone is not accurate enough to merit the authority of a serious "reliability" rating.

Project Description

The COVID-19 pandemic has been a paradigm shifting event for how the modern world conducts business, travel, and healthcare. The spread of the novel virus has been fast and brutal, and the limited knowledge of the virus during the first few crucial months have had massive ramifications. Information for the public is disseminated from authorities, hopefully, such as the CDC, the WHO, and politicians. Especially as the US government took steps to prevent the spread of COVID-19, the information from politicians to their constituents became more important. But what information was reliable? In the face of a public health crisis of such proportions, there were no options which would, in the end, satisfy everyone involved. Lockdowns have faced tremendous pushback, and not only by people who believe the virus is a hoax or over-hyped. People's livelihoods have been weighed against public health for the duration of the quarantine which originated as three weeks to #StopTheSpread. Now approximately eight months later, another lockdown seems to be on the horizon, but a vaccine might be as well.

More and more, public forum platforms such as Twitter are attempting to classify information based on reliability. The year of 2020 has seen wild links between the novel elements scrolling across timelines. For example, a simple wordbank algorithm seemed to have been checking for tweets containing the strings "5G" and "corona" to automatically cast doubt on the conspiratorial link between "5G" broadband and COVID-19. The outcome of this sort of automated classification is unclear in its prevention of the spread of disinformation. Currently, Twitter has been more aggressively tackling the spread of disinformation related to the 2020 election. These methods, however, have to either be manual or extremely computationally inexpensive given that they would have to scan practically the entire website. What if the scope of classifying reliability was narrowed while the process was taken to further depths?

This project is an attempt to classify the tweets of public officials for their reliability on spreading COVID-19 related information and give the associated politicians a "reliability score." While algorithms are inherently good at some things, something like language sentiment tends to be very complicated and extremely difficult to get right without very large training sets. Google Translate trains across millions of articles, speeches, books, and other forms of discourse and still does not quite hit the mark 100% of the time. The main hope for the project is that reliable and unreliable language will have shared traits that can be trained in a deep sentiment classification algorithm and be accurately extrapolated to other tweets. The algorithms can then classify about 100,000 text blurbs (as tweets) to give their author's a reliability score, which might have demographic predictors.

Dataset

For as many members of Congress (MOC) accounts[1] as possible, the most recent (up to one-thousand) tweets were collected using Twitter's API and put into individual handle CSV files. From here, using a COVID-19 word-bank, including everything from coronavirus to lockdown to CDC, the approximate 500,000 tweets were filtered down to around 100,000.

In order to fill out predictors for some visualization of the reliability data, another dataset was constructed using information compiled from multiple sources for the demographics of the members of the 116th congress and some other potentially interesting information. Though members were lost in every step of pre-processing for reasons such as non-

[1] Some notable accounts such as @AOC are absent, as only official accounts were used in this group.

existent or blank accounts, the final data set spanned 505 members of congress, or 94% of the 535 members.

Classifying a training set

For a training set passed to the algorithm testing for the sentiment of the tweets, two classifications were added to the COVID-19 related tweets. First, all of the tweets were placed into a large file stored with their corresponding handle and randomized. From these, 600 tweets were classified based on the COVID-19 relatedness and reliability. Being COVID-19 related was a simple binary classification as tweets which passed through the word-bank filter still may not be related. Secondly, each tweet was assigned a reliability score, but this was somewhat difficult, as most tweets do not simply tweet an easily known or proven fact. In order to better train the algorithm, the following sorts of classifications were used.

- $-\mathbf{1}-$ False information concerning COVID-19 in addition to using the virus as a political bludgeon[2]. Anybody using COVID-19 as an excuse for something else without any facts backing it would be unreliable.
- $\mathbf{0}-$ "Non-statement," tweets, from updating followers about COVID-19 related events, to discussion about the economic or social costs of the pandemic. Also included in this category are the many, many tweets thanking healthcare workers and the National Guard for their part played in the pandemic.
- $\mathbf{1}-$ Tweets containing factual information or helpful links. This is everything from giving health recommendations which are in line with the CDC guidelines, to sharing relevant statistics about the virus.

A changing definition of reliability

COVID-19 has been a massively changing phenomenon, and the information surrounding it has gotten better over time. Still, things did not start out perfectly. Early WHO and CDC information was off the mark, from speculating that the virus would not spread from person to person easily to suggesting that masks were not necessary. It can be hard to fault the members of congress for sharing early information such as this. Nevertheless, the standards for reliable and unreliable information are held to what is relevant today and for links shared at the time. This definition was upheld during the initial classification phase, but likely had little effect on the classification of the tweets, as discussed later.

Repairing and codifying the demographic dataset

While the dataset for senator handles was technically at the time of milestone II, there was still some work that needed doing with the dataset. For example, quite a few MOC had missing birthdays which needed to be added for age demographics. Additionally, some alignment problems with the previous scripts were solved manually. Some MOC have missing information concerning race/ethnicity, population represented, and median household income, and almost all of these were representatives from non-state territories like the Virgin Islands.

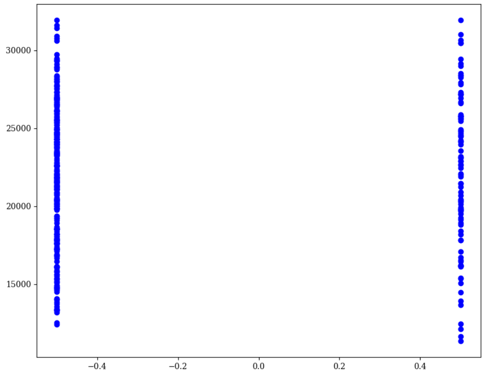

Figure 1: A graph of number-mapped gender against age as a number of days.

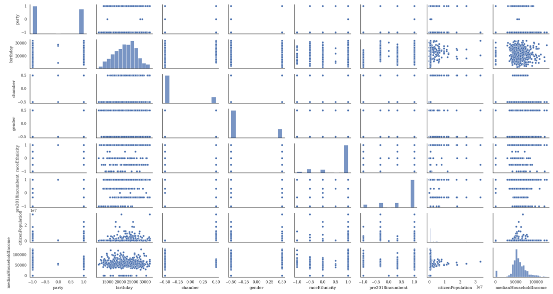

Figure 2: A seaborne pair-plot of the demographics under consideration for linear regression. Perhaps these demographics themselves tell a story.

For the regression algorithm, data concerning non-numerated classifications had to be codified into numbers. For example, the gender field had M correspond to -0.5 while F corresponds to 0.5. For a reference, see Fig. 1.[3].

This method of regression is mostly fine for binary data. Regression on data such as this when it comes to multiple classifications is interesting. "Pre2018incumbent" status, for example, was ordered from resounding incumbent wins to losses, with more neutral statuses for the seat such as not having an open race (in the case of the Senate) in the middle. Afterwards, there needs to be some analysis of the raw data for the purpose of prediction.

Several categories for demographics are highly dominated by one feature or another, such as the number of white MOC and the number of winning incumbents. More on the ramifi-

[2] Less than one out of 100 tweets are unreliable. Many representatives attempt to keep their timelines reliable. Therefore, the definition of unreliable was expanded to include blaming the virus or its spread on disputed causes.

[3] It was discovered afterwards that there are multiple ways to encode data such as this. Regression with multiple categorical input is very common, and the the dummy variables approach was chosen and made to have an unweighted mean of zero

cations of this in Evaluation.

Holes in the features A few members of Congress do not have current census data to back up their citizen population represented or median household income. For these MOC, blank values of 0 were used instead, which makes little difference on some scales such as citizen population represented where the relative difference between the many small districts and the massive is low. However, this percentage is easily visible for median household income where the mean is about 45,000. The solutions of giving placeholder values and removing incomplete MOC are both somewhat flawed, and perhaps the only way to truly deal with this is to train on the set of features sans those with placeholder values.

Approach

There are two algorithms and processes at play to move from 100,000 tweets to predicting reliability scores, the sentiment classification algorithm and the regression algorithm. The regression is simply multiple regression. The sentiment algorithm is the training of an NLP (Natural Language Processing) model. Language processing is highly complex, and as such two different approaches were tried and compared.

To clarify, the accuracy was based on the pre-classified tweets. For a test-train split of these human classified tweets, two different methods were used to read through the tweets and predict their sentiment. These were compared to the test set and validated for some sort of accuracy score. Due to the intense levels of volatility at every point along this curve, including: human curation, relatively small test sets, extremely subjective classification, difficult to classify tweets, several common threads between reliable and unreliable tweets (like all including #COVID19), and the volatility of the algorithms on such small datasets, the accuracy scores were highly inconsistent based on the test and training splits.

Linear Regression using Word Vectorizer

This simple Linear Regression model uses a Word Vectorizer found in Sklearn. The word vectorizer takes in all of the text from the tweets and converts each word into a vector that can be easily compared. This, in turn, allows us to extract features from the data. With this data, we can find which words appear more or less often in misleading tweets. This can be used to find potentially misleading tweets in the testing set based on their use of words that were commonly found in other misleading tweets in the training set. This algorithm was found to be exceptionally poorly performing.

Keras Sequential Model

The Keras sequential model is a neural network implemented in sklearn. This sequential neural net model is a linear set of layers. Each layer has a certain number of nodes as well, which is under the hood in these sorts of algorithms. To combat overfitting, the number of layers can be tuned for optimal validation accuracy (as training accuracy approaches 1.0 for deep enough networks). This algorithm saw much better performance and was used to classify the tweets.

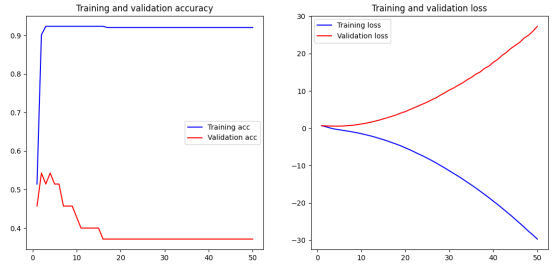

Figure 3: Initial scoring based on the tri-classification method. Very poor performance on the validation set due to the high conflation between the irrelevant and reliable tweets.

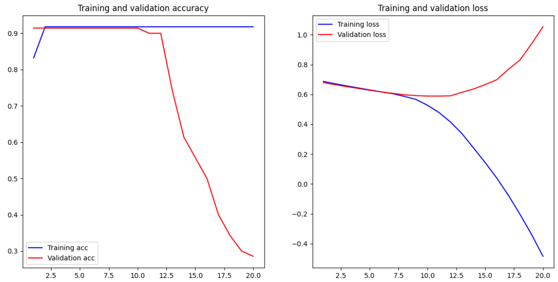

Figure 4: The sequential neural network performs much more strongly when it does not have to differentiate reliable and irrelevant tweets.

Evaluation

Tweet classifications, from ternary to binary

The Keras algorithm was trained across the 600 tweets, and came up initially with a very poor validation accuracy, see Fig. 3. This could be due to the sheer volume of data and **near indescribable difference between an irrelevant tweet (0) and reliable tweet (1)**.

Therefore, an executive decision was made to replace the 0 classification with a 1 for reliable, the machine learning equivalent of "innocent until proven guilty." The new model showed much improved results.

A few other hyper-parameters were chosen, such as updating from a countwords protocol to tokenized words, adding a bottleneck layer to squeeze the multiple parameters closer to the binary output, and the final best performing algorithm emerged with the following results. See Fig. 4.

Keras Sequential Model		
	accuracy	**loss**
training set	0.889	−0.823
test set	0.900	0.427

What follows are examples differently classified tweets. A **reliable** one:

> "It's critical that we listen to experts about how to reopen schools safely. Classroom instruction is important, but our kids (and teachers) need to be safe in the classroom.

> That can only happen if and when we ramp up testing and actually *have a plan*. Some ideas https://t.co/7u2rPJkyKl”
>
> - @RepKatiePorter 2020-07-11 16:23:28

And an **unreliable** one as designated by the network.

> “@luke_mo_ Nothing racist about where the virus came from (China), who lied about it (China), or who covered it up (China). If you read the story, you’ll be more informed about what they’ve done to your fellow Arkansans and Americans due to their actions.”
>
> - @RepRickCrawford 2020-03-18 21:01:21

Now, not all results can be so nice, and the ending split of about 17,000 unreliable to 95,000 seems skewed too harshly, but nevertheless, all tweets were classified, and their legislators thus scored. The score was a simple ratio of the reliable to unreliable tweets. Additional filters such as taking the square root of the ratio could have been used, but were ultimately deemed perhaps unnecessary. The slightly more generous split into unreliable tweets means that most members of Congress do not sit at 1.0 reliability. Most do stay close to the mean though, as the somewhat normal distributions for classifications in Fig. 5 shows.

Ultimately, nearly everyone had a few unreliable tweets, and most of these were likely the result of choosing words deemed unreliable by the ephemeral algorithm. Common blames in training such as China (who were blamed for deliberately releasing the virus) or the environment (where the PPM standards of the EPA have relation to how the virus spreads) were deemed unreliable in the training set, and therefore these common topics may have been deemed unreliable.

Additionally, while we filtered tweets automatically and manually for being COVID-19 related, **the tweets that the algorithm rated were only filtered automatically using a wordbank**. Therefore, some MOC likely were rated unreliable for talking about subjects irrelevant to COVID-19. An example of one such tweet is shown below.

> “ ‘For rural Montana, an accurate Census assures we get our fair share of federal funds for important priorities like infrastructure, education and health care.’
>
> Couldn’t have said it better, Paul! Complete your Census now at https://t.co/E8ylhzptXO.”
>
> - @SenatorTester 2020-09-08 17:55:49 (unreliable)

Demographic Predictors with Linear Regression

Regression score R^2 : 5-fold validation		
	mean R^2	$\pm$
With placeholders	-0.051	0.041
w/o placeholders	-0.056	0.028

A negative R^2 score means that the predicted weights perform worse than a simple horizontal line. Attempting to associate the the reliability of a member of Congress with some demographic score might actually hurt making an accurate prediction. The data is highly predicated on the test set used, and the scores range from about -0.25 to 0.10, with the score shown in the table being the cross-validated score.

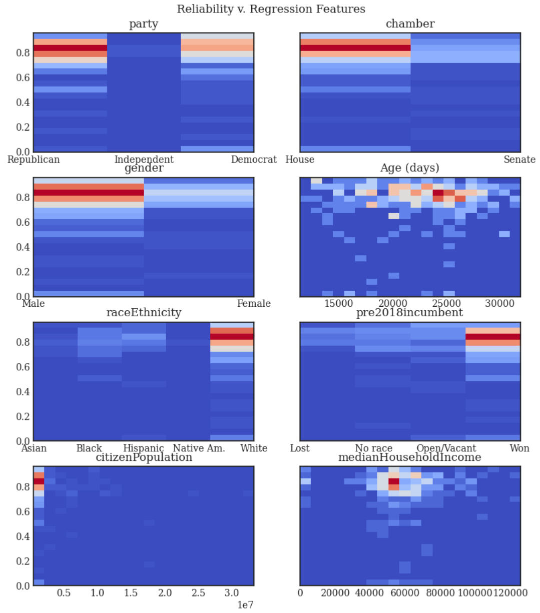

Figure 5: A heatmap of the predictor demographics. Note that due to a few highly concentrated groups, such was household incomes near the mean, white MOC, and winning incumbents, meaningful regression or correlation between demographics and reliability was highly hampered.

However, most people, regardless of any demographic predictors, hover around the mean of around 80% in terms of reliability, perhaps meaning that most people tend to normally disseminate what could be considered reliable (recall that the above approach consolidated reliable and irrelevant tweets into the reliable category). A horizontal line at about 80% is the best predictor, so the ultimate answer seems to be that, no, according to our tweet classifier, there are not significant demographic predictors of the reliability of COVID-19 information disseminated on Twitter from members of Congress.

Future work

Improving objectivity

The content of this project is by nature political and not immune to the inherent biases of the creators. When scoring the tweets, we attempted to sift through hundred of tweets of opinions and calls to action and objectively call them reliable and unreliable. Simply going on the offensive or casting blame for the virus on either side is not necessarily unreliable. Saying that the virus originated in China is not unreliable. Saying that President Trump knew of the virus in advance seems, according to multiple sources, to be a reliable take. The goal is not to make this project into an echo chamber for personal opinions, but certain things will most likely be classified differently for different people, and that seems unavoidable.

Using a legal sort of argumentative basis can help with defining objective reliability. To say that President Trump knew in January that the virus was more deadly than the flu seems like a corroborated fact, but to say that he is directly killing the American people with his actions are harder to prove. If the case could not be proven in court, perhaps the claim is unreliable.

As mentioned previously, inherent bias is unavoidable anywhere and especially comes to play in this manual classification of the training data. In order to remain as objective as possible when deeming tweets as either reliable or unreliable it was important for team members to be cognizant of any biases we may have. While unavoidable, when classifying in the spreadsheet, the column displaying the names of the MOC was hidden in an attempt to mitigate as much bias as possible. When it came to actually determining whether a fact contained in a tweet was true or not, research was done and multiple sources were consulted to ensure the accuracy of classification. Most tweets deemed reliable contained relatively common sense facts such as ”wash your hands” or ”wear a mask” which made further research unnecessary. When it came to unreliable tweets, however, much deeper research was needed. For example, the following tweet was deemed unreliable:

> “ ‘In the middle of a pandemic, the Trump administration raised the application fee to become a U.S. citizen from $640 to $1,160. In 1995, it was $225. This is extortion and part of the Trump administration’s war on immigrants – regardless of legal status. Citizenship4All https://t.co/OXPFI2ljho”
>
> - @RepVeasey 2020-09-17 20:07:55 (unreliable)

While it is factual that the cost of filing the N-400 Application for Naturalization form will increase to $1,160, the decision was made by the U.S. Citizenship and Immigration Service, an agency within the Department of Homeland Security. The sentiment of the tweet is clearly attempting to place blame on the Trump administration for this change, and while the President does have ultimate authority over government agencies and their personnel, there is no evidence that this change came directly from him or his direct subordinates. Additionally, the tweet attempts to convey that the change was made during the pandemic. While this change did take effect on October 2nd, according to the 576 page official document outlining fees and scheduling and published by the Department of Homeland Security, the fee changes for the 2020 fiscal year were decided in the 2017 fiscal year, long before the emergence of the pandemic. Due to these misleading sentiments, it was deemed unreliable. This is just one example of the extensive research that went into verifying the validity of claims made by MOC in their tweets in the training data.

Increasing the size of the training data

This project was an ambitious undertaking, as it not only sought to codify and predicatively score discourse, but sought to do so on a relatively massive scale for four undergraduates. The project required some grunt work in the pre-processing phase, and the very imprecise nature of this classification meant that the data was to be expected to be very noisy. If anything, this project needed a massive amount of pre-processing, I.E. more people rating more tweets. This sort of algorithm could roll out on Twitter, and some variations of an algorithm like this have, though they likely used much more simplistic wordbanks (such as ”5G corona” on Twitter).

In the given time frame, the four team members cumulatively were able to classify around 600 tweets for reliability. For use in deep learning algorithms, this is an absurdly low amount of training data. It is very likely that had more training data been fed into the algorithm and it was tweaked as more data was read in, the results could have been much better as the algorithm ”learned”. As previously mentioned, however, it was not within the scope nor time frame of this project to classify additional tweets.

Another important note relates back to the objectivity of the manual classification. With only four members classifying tweets as reliable or unreliable, the data is much more prone to the biases of the classifiers. If this task could have been delegated to a large number of people each classifying fewer tweets, the biases would have likely made less of an impact. Unfortunately this was not possible and the effects of biases not easily measurable and so the project continued on.

Additionally, this sort of algorithm tends to be quite callous and unfair to those about whom it tests. I certainly would not enjoy if Twitter rolled out an algorithm which rated my on my reliability decided by a machine, and some users are hurt more than others. People with few tweets easily fell through the cracks in our approach.

Simple, binary reliability

A question dodged since beginning this project was to define truth and reliability and how to make a machine capable of classifying based on this definition. Perhaps classifying over 100,000 tweets on reliability is too broad in scope. Politicians are professional communicators representing their constituents and maintaining public favor and confidence. They are not scientific journals tweets cold, unfeeling facts. Therefore, when it comes to assessing the truthfulness of their tweets, much if not most of the work is down to personal opinion.

Most everything tweeted by members of Congress is arguably true, and typically represents an opinion of some kind. Saying that Americans need a COVID-19 relief bill is true in nearly every respect but was normally classified as $\mathbf{0}$, or irrelevant, given that this tweet does not make some statement about the virus itself. Tweets about transmission vectors or help in preventing the spread were considered for a $\mathbf{1}$ in reliability as they were typically relevant and true, but this can backfire as well. The problem is that politicians are not health boards, their job is not tweeting out dry facts. The task of handling all of the information related to COVID-19 and boiling thousands of opinion pieces covering a multitude of subjects down to a single reliability score may have

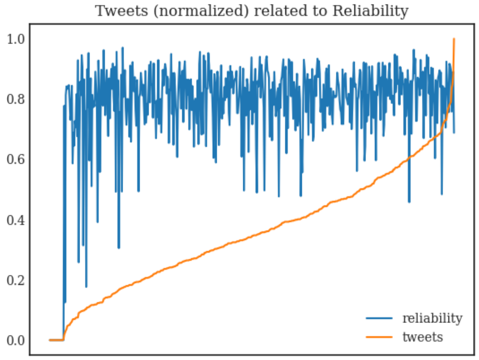

Figure 6: Graph of the correlation between reliability score and number of tweets relative to the rest of the members of congress. Note that for accounts with zero tweets, a default score of zero was given. There is a slight correlation between having relatively few COVID-19 related tweets and a low score, given that more tweets discussing other issues might have been flagged as unreliable, and a small sample size makes unreliable tweets more heavily weighted.

perhaps been too broad of a scope.

A balance was carefully considered between over-classifying negative examples for more data and reserving the tag for only the most overt lies. Not every doomsayer or clear sky politician is being inherently unreliable based on misinformation. When a tweet says, "we're doing everything we can to fight COVID-19," different people would have different opinions as to whether that is reliable, especially considering on when the tweet was sent, a sentiment NOT shared by the sentiment algorithm.

The Robotic Authority

Hesitation and discretion are encouraged approaching the idea of some algorithm which classifies tweets without several people reading them to score someone on their reliability. Tokens of reputation handed down from higher authorities that did not have significant human curation are not reliable, and the reliability neural network algorithm, though to me highly interesting and impressive, should not be adhered to as some voice of objective authority. There are upstream inefficiencies and biases in the process, and while a few of the examples of picking out reliable and unreliable tweets exist, several more seem to have somewhat arbitrary placement.

> "US Capitol coronavirus COVID19 update by CDC FDA NIH
>
> *90% of all USA drug content comes from China
>
> *USA challenge is still low risk
>
> *NO USA quarantine helpers have COVID19
>
> *USA testing kits more available
>
> *USA has PERFECT zero fatality record
>
> *COVID19 NOT man-made https://t.co/Y4QZNM1GEr""
>
> - @RepMoBrooks, 2020-02-28 13:36:01 (reliable, but perhaps at the time it was)

Sentiment matching can be gamed by simply speaking with authority in a somewhat objective tone. Any tweet that lists off COVID-19 facts backed with numbers, perhaps however wrong, stands a chance to be marked reliable. An algorithm can be highly reliable, but for something as complicated as language, we are far away from trusting it to disseminate true and untrue information, especially without consulting some databank. Otherwise, the model is an image of its programmers.

Team Member Contributions

The paper was nearly entirely imported from Milestone 3, with a few changes for formatting and wording to adhere to AAAI formatting. For the presentation, everyone did their share of the slides and presented them well at the final project presentations. Thanks for a great semester!