Machine Learning: The Accuracies and Shortfalls of Artificial Intelligence in Fact Checking Presidential Candidates

A full-length DSCI 133 report authored by Rishik Hombal, Brandon Rudolph, and Jack Zhang
rsh83@case.edu
brr30@case.edu
cxz416@case.edu

The following students have contributed equally and should be awarded equal credit:

Rishik Hombal

Brandon Rudolph

Jack Zhang

Progress, Timeline, and Responsibilities:

March 7th: Requested approval from Professor Li for end of year project. Group consisting of Rishik Hombal, Brandon Rudolph, and Jack Zhang acquired approval do research on Machine Learning methods used to detect false news statements and/or fake news stories.

March 8th-17th: CWRU Spring Break, no group meetings.

March 18th: Group meeting to discuss project ideas and general direction of report. At this point we decided to focus our research on fact-checking based machine learning algorithms. This decision is preceded by extensive research into Fake News article detection; we decided data would be easier to collect, train, and test through usage of individual strings of text as opposed to entire online articles.

March 19th-23rd: Research into possible sources for data collection.

March 24th: Group meeting to discuss possible data sources. We then conclude that Twitter's API would be an adequate resource to use considering its openness, accessibility, scope, and vastness. We then decided our eventual code would be tested on rising Presidential candidates due to their place in the public eye and tendencies to make statements with higher chances of containing false information.

March 25th-31st: Data collection process commenced. Further research into possible AI code.

April 1st-11th: Heavy exam schedules, enough progress made at this point to justify priorities.

April 12th-14th/April 19th-21st: Report authored, code developed and tested, and results evaluated as project final steps are completed.

April 25th: Report findings presented to DSCI 133 class.

April 29th: Last day of classes. Project submitted by this deadline.

Responsibilities: All group members unanimously agree that an equal amount of time and effort was spent by each member in completion of this report and that equal grades should be awarded as a result. Rishik headed data collection and cleaning, Jack headed development of machine learning usage, and Brandon headed analysis of data and report structuring.

Machine Learning: The Accuracy and Shortfalls of Artificial Intelligence in Fact Checking Presidential Candidates

An academic paper authored by: Rishik Hombal, Brandon Rudolph, and Jack Zhang

Department of Electrical Engineering and Computer Science, Case Western Reserve University

Cleveland, OH 44016-7076

April 29, 2019

Abstract:

The purpose of this report was to train a machine learning algorithm to fact check Presidential candidates. Using training data provided by William Yang Wang, we trained Google's machine learning libraries Keras and Tensorflow to develop a supervised-learning algorithm to determine the truth of the program's inputs. Outputs are enumerated from 0 (not true) through 5 (entirely true), spanning a set of 6 possible results per input. After testing the algorithm, it received an accuracy of 30%. We then proceeded to use our data to make correlational connections between the honesty of a candidate and their political performances.

I. Introduction:

The invention of information technology, specifically devices created for day to day usage, has revolutionized methods and rates of communication. The development of smartphones and social media has created an interconnected virtual world where data can be shared and manipulated by anyone. One need not look further than the controversial and continuously discussed 2016 American Presidential Election as an unfortunate, yet predictable shortfall of the recently conceived informational technological world.

Even so, the importance of new forms of communication cannot be ignored and have changed the way politicians interact with potential voters. The current President, Donald Trump, has almost eliminated White House Press Briefings in favor of Twitter. In fact, the current administration did not have a Press Briefing for a record 6 weeks from December 18th, 2018 to March 11, 2019 [2]. In that time, the government shut down for a period of 35 days, both the White House Chief of Staff and the Secretary of Defense left office, US troops were withdrawn from Syria, among many other significant stories [3]. These events, however, were accompanied by a Tweet from the President – some of these Tweets being the first announcements of these events to the public. This shift from press briefings to Tweets signals a shift from press-focused media distribution to people-focused media distribution. Regardless of the politics surrounding that shift, it certainly has been effective: a 2018 Gallup poll found that "76% of Americans see, read or hear about Trump's tweets" [4]. This success in sheer distribution has led to candidates for the 2020 Presidential Election to invest heavily in a social media presence. All 2020 candidates have a Twitter account that tweets important campaign information and news. Joe Biden, one of the expected Democratic frontrunners, after being accused of inappropriate conduct, released a video defending himself on his Twitter before making any formal press release. Thus, to truly gain insights on the statements made by candidates, one must turn to their Twitter - the collection of a large portion of their policy.

In order to decide whether information is true, we must first set a standard for what truth is. Propositional logic is concerned with determining whether a proposition, or some type of combination of such propositions, holds a truth value: true or not true. In binary this can be expressed through 0 (false) and 1 (true). However, not all sentences contain such propositions which can be clearly quantified as true or false. Null, or unknown, is the third truth value which can be used in our quest to fact check statements. Something which is unknown could be either true or false [1].

Truth can also be quantified. The concept behind a confidence interval is predicting the value of an entire population based upon the results of a sample set. Polling is an excellent example of this concept. When a poll is run through an institution such as Gallup, data on a subset of an entire population is gathered. From this subset, statisticians attempt to draw conclusions on that population, accompanied by a confidence interval to account for margin of error. Even that data, however, is rarely concerned with the truthfulness of something, but instead upon opinions – something hard to scientifically quantify outside the scope of that particular poll. The aim of this project is to establish a means of rating candidates according to an absolute scale of truthfulness so voters can be more informed when they make a decision in the overcrowded 2020 primary elections and then the Presidential election.

II. Methods

As of this moment 21 candidates have announced or are extremely likely to announce plans to run for president in 2020: 19 Democrats and 2 Republicans. With the Iowa caucuses being nearly nine months away, we could even expect a few more people to throw their hat into the ring. All 21 of these candidates have official twitter accounts which they use to communicate with the public. As noted in the introduction, social media can often be a more effective tool at spreading information compared to TV interviews or press conferences, making it a strategic necessity for a candidate for public office. On the social media platform of Twitter there are, on average, 500 million tweets per day [5]. Due to comprehensive usage of Twitter by the candidates and the open-to-public nature of each candidates' handle, Twitter was deemed to be the best resource of data to test our algorithm on.

The Data:

To test our machine learning algorithm, we will assemble our own dataset consisting of the tweets which these 21 candidates have published. To gather this data, we first acquired approval from Twitter to use Twitter Developer's API. There was a wait time between the point which we requested access and gained access: the purpose of this overlay was cited as a security concern. "Twitter also emphasized it was publicizing its data so third-party research and analysis could be made. 'We are making this data available with the goal of encouraging open research and investigation of these behaviors from researchers and academics around the world. This is the continuation of our overarching mission to serve the public conversation'" [6]. Although Twitter bots is not the exact subject of the report, they are a contributing factor to the extensive research and public debate over fact-checking algorithms. The purpose of the extensive security protecting Twitter's platform happens to be a cause of our research into AI fact checking: if Twitter can accurately recognize misleading statements, it has a better chance of removing potentially harmful viral accounts before they become apart of the community as a whole. Following approval from Twitter Development, we received Customer and Access keys, passwords which allowed our code based around Tweepy to operate.

We then developed code using Python 3.0 and the Tweepy package. Using Tweepy, we gathered the 200 most recent tweets, the maximum allowed by Tweepy per call, from each candidate before that number was reduced by data cleaning. To clean the data, we first eliminated retweets. Retweets, while sometimes important, were usually links to articles or campaign pictures, and sometimes just images or videos which did not have usable text. To further clean the data, we eliminated non-Ascii characters using the encode and decode function. We also removed symbols such as hashtags, mentions, dollar signs, and other non-English characters using the replace function because we suspected that those characters could lower the accuracy of our neural network testing because they rarely appeared in our training set. The Tweepy package also returned the link to the tweet in addition to the tweet itself so those links were also removed using the "rfind" command. In all, from all 21 candidates, we gathered over 3250 usable tweets. The data was exported as XLSX files - Excel workbook files. We originally attempted to export data as CSV files, however, the fact that the tweets themselves would contain commas essential to the understanding of the tweet meant we could not just remove commas. These XLSX files contained a single column containing the tweet and the files were named with the twitter handle of the candidate. These files were then used to test our machine learning algorithm.

Building he Algorithm:

To train our algorithm, we turned to existing fact checking data sets such as the one created by

William Yang Wang [11]. Wang used the now defunct Politifact API to fact check a series of political statements made on social media, television, radio, and other mediums and assign a truthfulness score ranging from "Pants on Fire", the most untrue statement, to "true", a completely true statement. Wang's dataset is the most extensive dataset analyzing the truthfulness of claims and contains over 12.8 thousand entries. Using techniques we discussed in class and practiced in our lab assignments, we manipulated the data to use on our machine learning algorithm. The first thing we did was to read Wang's data into python as a Dataframe. We split this data into two columns: "Validity" and "Comment". Validity represented the truthfulness, we took Wang's scoring of "pants-on-fire", "false", "barely-true", "mostly-true", and "true" into a number 0-5 and set this as the y-training set. The actual comment as a string which we converted into TF-IDF(term frequency–inverse document frequency), which is a numerical representation of the sentences, and set this equal to the x-training set. These two training sets were then fed to our neural network. For the algorithm, we developed a pretty simple neural network model using Google's machine learning libraries Keras and Tensorflow [19]. The neural network was simple containing only two hidden layers as opposed to more complex ones that can have up to thousands. We used a "cross-entropy loss function" which handles multiple outputs (0-5) with Adam as our optimizer, an improved version of the Gradient Descent discussed and practiced in DSCI 133. We split our input data into 90% to let our algorithm learn on and used the 10% as a validation set to see how accurate our model was. Compiling the model over 5 epochs with a batch-size of 200, we got model with a 30% accuracy. For perspective, the much more complex and cutting-edge text analyzing models from MIT only have an accuracy of 65% [7]. We then ran our model on the data of tweets for each candidate, and saved their truth scores into excel files which we used to create our graphs and analyze the data.

III. Results:

After gathering the XLSX files, the data had to uploaded to Jupyter Notebook, an open web application which allows one to use python for data visualization. In Jupyter, we read each file corresponding to each of the 21 twitter pages analyzed. Using the Pandas software library, we loaded each set of data into a single Dataframe. To get a better sense of what types of information we had collected, Table 1 consists of the basic statistics for each of the candidates based upon the algorithm used.

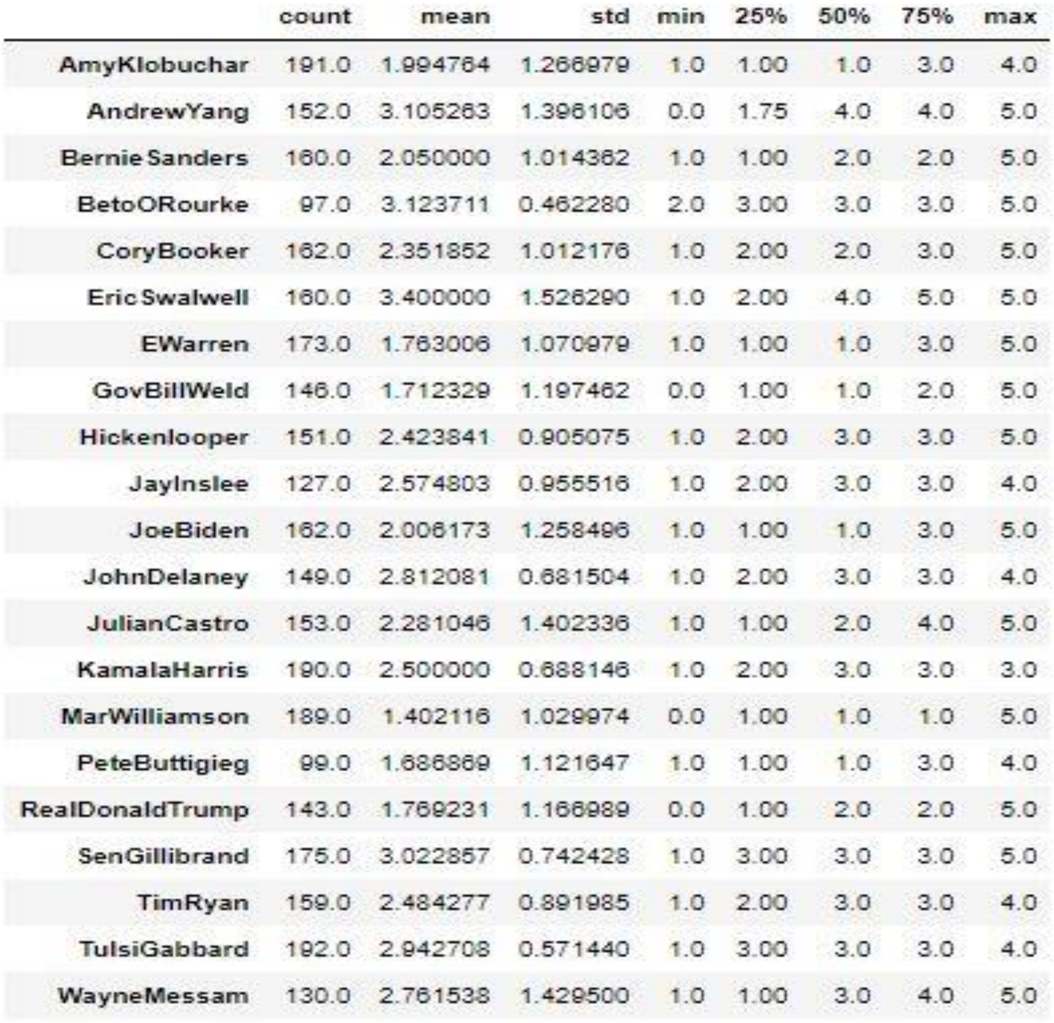
Next, two bar graphs were created to allow visualization of the mean scores, sorted from lowest to highest, and of the standard deviations, also sorted from lowest to highest. Figure 1 allows us to get a better sense of how each candidate faired both their consistency and honesty.

Figure 2 consists of 21 distributions for each candidate analyzed. This allows us to visualize the disparities of each candidate in how they tweet in addition to how their data might be skewed.

Figures 1 and 2 only tells us, according to our testing, who is more trustworthy compared to peers on social media. While these figures are important to consider when listening to how a politician communicates with voters, we wanted to see if there were any types of predictions, we could make about the candidates based upon these values. The first variable we wanted to test was public opinion. Public opinion does not necessarily tell which candidates are more popular within an election, but instead it measures how the public views a single politician alone. The logic behind this approach was to see if there was a positive correlation between candidates' truth scores and their public opinion: the more honest the person, the more favorable view the public has of them. Unfortunately, there were no comprehensive polling data to test this hypothesis. The next variable which we decided to test was first quarter funding numbers [8]. The thought process driving this test was all the headlines lately involving campaign finance laws and regulations. We hypothesize that there would be a negative correlation between the honesty (truth values) and funding (in millions) since the less truthful a politician was, the more donors would be willing to finance their campaign. Although our linear regression slope was negative, the model's score was too low to conclude any direct correlation between the two variables. This regression graph can be viewed in figure 3.

Next, we sought to collect polling data for likely voters of the 2020 democratic primary due to its wide availability. We used data from fivethrityeight.com [9] which asked residences of Iowa, the first state to vote in the primary election, whom they planned to caucus for. A handful of original handles we collected tweets from and analyzed did not appear anywhere on these polls because they were either Republican or not popular enough to register. The data was cleaned of these people as a result and we then preformed linear regression test on polling vs funding. The thought process behind this method was that the success and the finances of a candidate would be positively related. This was true as there was a positive regression slope and a model score of .76. From here, we wanted to factor in our own data to see if we could improve the performance of the model by also weighing a third variable. We decided to plot polling vs the ratio of funding over truth value. The ratio is to be used as a way of determining the authenticity of each dollar raised, meaning a higher score would correspond with a less honest and better funded politician. When we take this regression, the accuracy of our model actually improved, showing a correlation model score of .85, .09 points higher than of polling vs funding alone. The scatterplots, regression lines, and model scores from our python code can be seen in figure 4.

Table 1



	count	mean	std	min	25%	50%	75%	max
AmyKlobuchar	191.0	1.994764	1.266979	1.0	1.00	1.0	3.0	4.0
AndrewYang	152.0	3.105263	1.396106	0.0	1.75	4.0	4.0	5.0
Bernie Sanders	160.0	2.050000	1.014362	1.0	1.00	2.0	2.0	5.0
BetoORourke	97.0	3.123711	0.462280	2.0	3.00	3.0	3.0	5.0
CoryBooker	162.0	2.351852	1.012176	1.0	2.00	2.0	3.0	5.0
Eric Swalwell	160.0	3.400000	1.526290	1.0	2.00	4.0	5.0	5.0
EWarren	173.0	1.763006	1.070979	1.0	1.00	1.0	3.0	5.0
GovBillWeld	146.0	1.712329	1.197462	0.0	1.00	1.0	2.0	5.0
Hickenlooper	151.0	2.423841	0.905075	1.0	2.00	3.0	3.0	5.0
JayInslee	127.0	2.574803	0.955516	1.0	2.00	3.0	3.0	4.0
JoeBiden	162.0	2.006173	1.258496	1.0	1.00	1.0	3.0	5.0
JohnDelaney	149.0	2.812081	0.681504	1.0	2.00	3.0	3.0	4.0
JulianCastro	153.0	2.281046	1.402336	1.0	1.00	2.0	4.0	5.0
KamalaHarris	190.0	2.500000	0.688146	1.0	2.00	3.0	3.0	3.0
MarWilliamson	189.0	1.402116	1.029974	0.0	1.00	1.0	1.0	5.0
PeteButtigieg	99.0	1.686869	1.121647	1.0	1.00	1.0	3.0	4.0
RealDonaldTrump	143.0	1.769231	1.166989	0.0	1.00	2.0	2.0	5.0
SenGillibrand	175.0	3.022857	0.742428	1.0	3.00	3.0	3.0	5.0
TimRyan	159.0	2.484277	0.891985	1.0	2.00	3.0	3.0	4.0
TulsiGabbard	192.0	2.942708	0.571440	1.0	3.00	3.0	3.0	4.0
WayneMessam	130.0	2.761538	1.429500	1.0	1.00	3.0	4.0	5.0

Figure 1

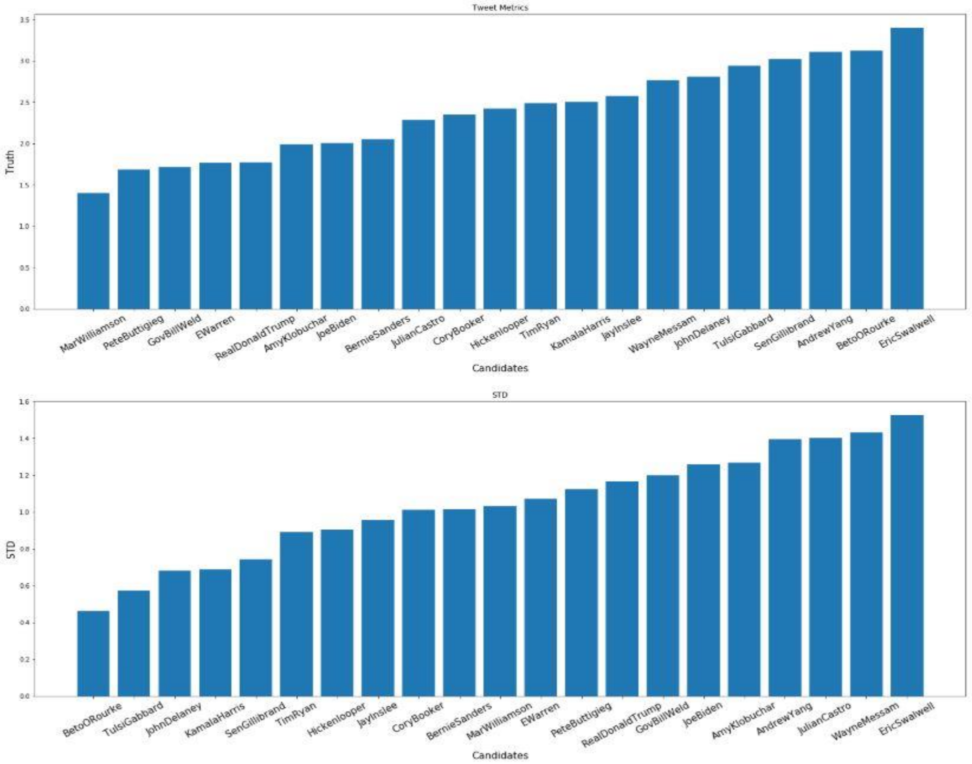


Figure 2

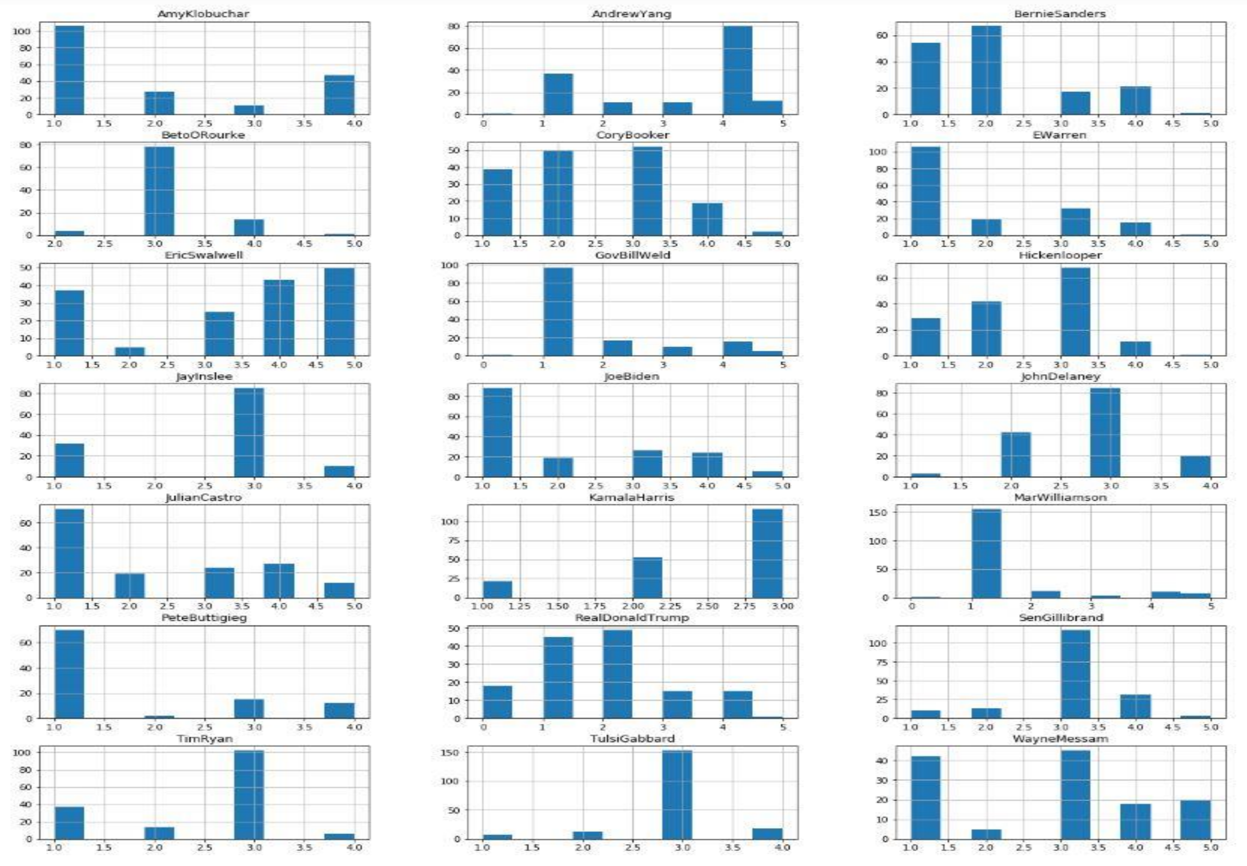
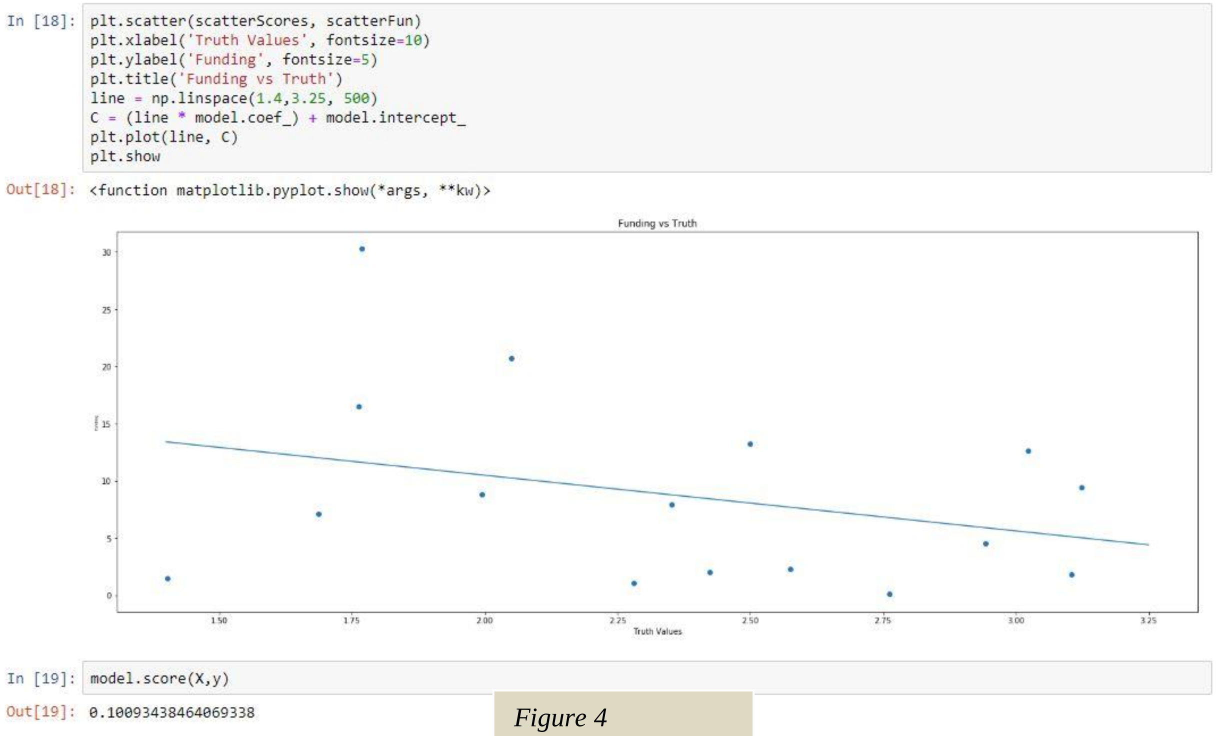


Figure 3



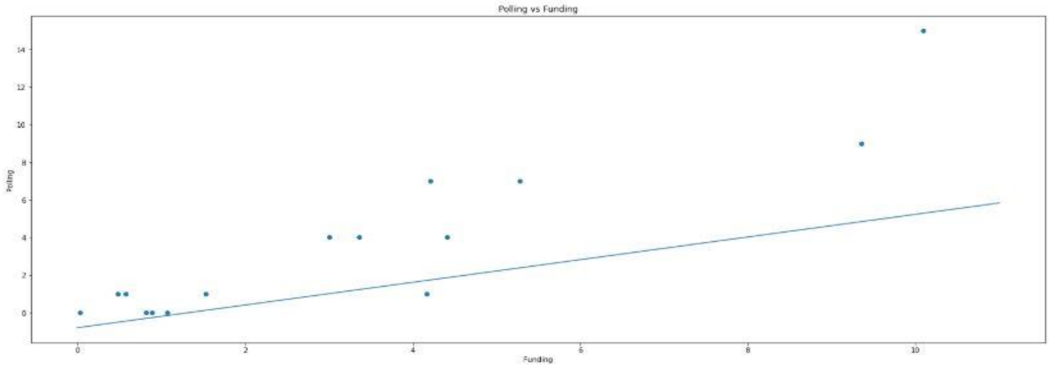
```
In [18]: plt.scatter(scatterScores, scatterFun)
         plt.xlabel('Truth Values', fontsize=10)
         plt.ylabel('Funding', fontsize=5)
         plt.title('Funding vs Truth')
         line = np.linspace(1.4,3.25, 500)
         C = (line * model.coef_) + model.intercept_
         plt.plot(line, C)
         plt.show

Out[18]: <function matplotlib.pyplot.show(*args, **kw)>
```

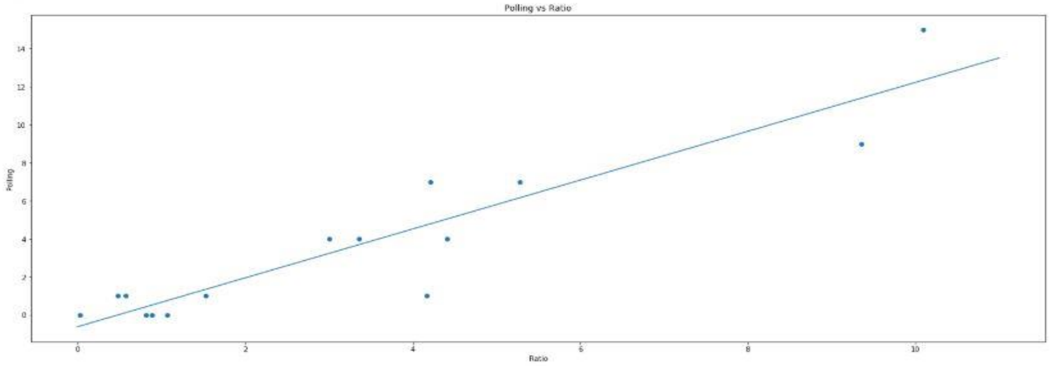
```
In [19]: model.score(X,y)

Out[19]: 0.10093438464069338
```

Figure 4



```
R Squared of  0.7638249941458373
```



```
R Squared of  0.8571993106393145
```

Analysis:

In DSCI 133, we have discussed a handful of ways to analyze the data we gather with the intention of discovering underlying trends. In this section, we will discuss not only the what, but the why: are there other variables which can be traced back to the quality of trustworthiness and truth telling within a particular political candidate?

Fake News and widespread misinformation is a relatively new phenomenon. The nature of technology and social media dictates that the constant change to our means of communication only make regulation of the dangers associated with it more difficult to develop. Although top academic institutions have been concerned with the problem of fake news and fact checking, a considerably low amount of progress has been made in the development of software to detect faulty news statements. Karen Hao of the MIT Technology Review admits that, “The best model accurately labeled news outlets with low, medium, or high factuality just 65% of the time” [7]. The fact of the matter is that this is a new field of research. Social media platforms such as Facebook and Twitter are hardly a decade and a half old and have only begun to rise into its overwhelming prominence over the last few years. The current President uses Twitter as his primary source of communication with the American public; President W. Bush, who left office in 2009, didn’t even have a Twitter account. The technology is progressing at a rate that is almost too difficult to keep up with.

As with the technology that we’ve spent this semester researching, the very nature of facts can change. Lucas Graves of Oxford University expands upon the limits to constantly changing basis data. He writes, “In practice, fully automatic verification today remains limited to experiments focused on a very narrow universe of mostly statistical claims. For instance, both Argentina’s Chequeado and the UK’s Full Fact are developing purpose-built AFC modules designed to match claims about specific public statistics, such as the unemployment or inflation rate, against the official figures” [13]. Another issue with fact-checking technology is that there is no single collection of training data which can be used. As the events in the world occur and new discoveries are made, the data which is to be considered the standard for truth most also change. That highlights one of the shortfalls of our algorithm: the age of our training data. We chose to use William Yang Wang’s data set primarily because it was the only such dataset with any level of comprehensiveness, unfortunately, due to the age of the dataset, some more recent facts were unable to be reliably fact checked by our algorithm.

In addition to being a relatively new issue, determining the level of truth in a statement is an awfully difficult job to perform. Brooke Borel, a journalist from Popular Science expands upon the complexities of truth. She writes, “[T]ruth is complex and squishy. Human brains can recognize context and nuance, which are both key in verifying information. We can spot sarcasm. We know irony. We understand that syntax can shift even while the basic message remains” [10]. As discussed in the Introduction to this paper, the very definition of truth can take on numerous meanings and interpretations.

IV. Conclusion:

In setting out to complete this project, we gave ourselves a tremendous task: to do with a machine what some reporters have dedicated their entire lives to. As such, our results were not ideal. Our accuracy was low, and we ended up creating even more questions. A lot of this can be chalked up to our skill level – we as freshman are not properly equipped to handle a problem which has befuddled respected and accomplished professionals for years. Even with human discretion, there are issues navigating the gray areas as to whether something is truthful. This made our task even more difficult. Machine learning and Artificial Intelligence faces significant shortcomings in this area and our algorithm faced even more shortcomings due to our skill level and the data we had access to; however, I believe the work we did is an instrumental first step in this field. There are no other projects which have been fact checking the Twitter feeds of Presidential candidates on this scale. Instead smaller projects have chosen to fact check a small subset of tweets of a particular candidate or a few candidates by hand. That method falls prey to subjectivity. Instead, we have easily reproducible results which should only continue to improve as more training data is added. This project is not perfect by any means, but we have produced novel work which is pioneering in this field. We only hope that the data we have gathered can assist voters in making an informed and qualified decision in the 2020 Presidential Election.

Acknowledgments:
We would like to thank Dr. Pan Li, Tianxi Ji, and the CWRU Department of Electrical Engineering and Computer Science in their guidance toward the completion of this report and its discoveries.

References:

Source links:

[1] Tutorialspoint.com, “Discrete Mathematics Propositional Logic,” *www.tutorialspoint.com*. [Online]. Available: https://www.tutorialspoint.com/discrete_mathematics/discrete_mathematics_propositional_logic.htm. [Accessed: 18 Apr. 2019].

[2] Mallin, A. (2019). *White House sets record for time span with no press briefings.* [online] ABC News. Available at: https://abcnews.go.com/Politics/white-house-sets-record-longest-span-press-briefings/story?id=60472803 [Accessed 26 Mar. 2019].

[3] Higgins, T. and Breuninger, K. (2019). *White House announces press briefing after record six weeks without one*. [online] CNBC. Available at: https://www.cnbc.com/2019/03/11/white-house-announces-press-briefing-after-record-time-without-one.html [Accessed 2 Apr. 2019].

[4] Newport, F. (2018). *Deconstruction Trump's Use of Twitter*. [online] gallup.com. Available at: https://news.gallup.com/poll/234509/deconstructing-trump-twitter.aspx [Accessed 11 Apr. 2019].

[5] Internetlivestats.com. (2019). *Twitter Usage Statistics - Internet Live Stats*. [online] Available at: https://www.internetlivestats.com/twitter-statistics/ [Accessed 26 Apr. 2019].

[6] Romano, A. (2018). *Twitter released 9 million tweets from one Russian troll farm. Here’s what we learned.*. [online] Vox.com. Available at: https://www.vox.com/2018/10/19/17990946/twitter-russian-trolls-bots-election-tampering [Accessed 5 Apr. 2019].

[7] Hao, K. (2018). *AI is still terrible at spotting fake news*. [online] MIT Technology Review. Available at: https://www.technologyreview.com/s/612236/even-the-best-ai-for-spotting-fake-news-is-still-terrible/ [Accessed 21 Apr. 2019].

[8] Axios.com. (2019). *2020 presidential election: Track every candidate's Q1 fundraising totals*. [online] Available at: https://www.axios.com/2020-presidential-election-campaign-fundraising-tracker-e37049f7-f596-40d5-a7d2-bff3d3940d86.html [Accessed 20 Apr. 2019].

[9] FiveThirtyEight.com. (2019). *Iowa’s 2020 Democratic Presidential Caucuses*. [online] Available at: https://projects.fivethirtyeight.com/2020-primaries/democratic/iowa/ [Accessed 24 Apr. 2019].

[10] Borel, B. (2018). *Consent Form | Popular Science*. [online] Popsci.com. Available at: https://www.popsci.com/can-artificial-intelligence-solve-internets-fake-news-problem [Accessed 16 Apr. 2019].

[11] Wang, W. (2017). *"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection*. [online] arXiv.org. Available at: https://arxiv.org/abs/1705.00648 [Accessed 30 Mar. 2019].

[12] Skelley, G. (2019). *What Can A Candidate’s Home-State Popularity Tell Us About 2020?*. [online] FiveThirtyEight.com. Available at: https://fivethirtyeight.com/features/how-the-

2020-candidates-do-with-the-voters-who-know-them-best/ [Accessed 15 Apr. 2019].

[13] L. Graves, Reutersinstitute.politics.ox.ac.uk. (2019). [Online] Available https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/graves_factsheet_180226%20FINAL.pdf [Accessed 23 Apr. 2019].

[14] Mello and Fernandes (2018). *EBSCOhost - world's foremost premium research database service*. [online] Search.ebscohost.com. Available at: http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=cat07006a&AN=cwru.b5705217&site=eds-live [Accessed 03 Apr. 2019].

[15] Singer-Vine, J. (2016). *BuzzFeedNews/2016-10-facebook-fact-check*. [online] GitHub. Available at: https://github.com/BuzzFeedNews/2016-10-facebook-fact-check [Accessed 13 Apr. 2019].

[16] Mello, R. F. de, & Ponti, M. A. (2018). Machine learning : a practical approach on the statistical learning theory. Springer. [online] Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=cat07006a&AN=cwru.b5705217&site=eds-live craigsilverman/partisan-fb-pages-analysis [Accessed 7 Apr. 2019].

[19] TensorFlow. (2019). Keras | TensorFlow Core 2.0a | TensorFlow. [online] Available at:https://www.tensorflow.org/alpha/guide/keras [Accessed 14 Apr. 2019].

Dataset:

[11] Wang, W. (2017). *"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection*. [online] arXiv.org. Available at: https://arxiv.org/abs/1705.00648 [Accessed 30 Mar. 2019].

Major References:

[17] Shu, Kai & Sliva, Amy & Wang, Suhang & Tang, Jiliang & Liu, Huan. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter. 19. 10.1145/3137597.3137600. [Accessed 15 Mar. 2019]

[18] Davis, M. (2016). *Hyperpartisan Facebook Pages*. [online] Buzzfeednews.com. Available at: https://www.buzzfeednews.com/article/