Utilizing Learning Algorithms to Improve News Transparency and Electoral Predictions

Asim Hirji

HMN 679HB

Departmental Honors in the Humanities Program

The University of Texas at Austin

May 2020

_____

Prof. Brian E. Roberts
Department of Government
Supervising Professor

_____

Prof. Christopher Wlezien
Department of Government
Second Reader

Table of Contents

1. Abstract

This paper investigates how the increased volume of data in our political system can transform the role of data in politics itself. To explain further, data are involved in many different aspects of politics including polling, candidate approval ratings, and even the click through rate of a candidate website. This research, however, focuses on two main themes: 1) using data to improve on identifying whether what political candidates say is true or false, thereby improving the spread of what people and the media consider credible news, and, 2) using different types of data to predict election outcomes. More specifically, in the context of predicting elections, this research focuses on the role social media can play in election forecasts, the role of polling data regarding particular candidates for election forecasts, and specifically the role of polling data regarding issues such as the status of the economy, unemployment rate, etc. for election forecasts. To perform these tasks, this research takes both a qualitative and quantitative approach when observing data. From these objectives, this paper utilizes quantitative analysis of different variables feeding into these objectives and utilizes machine learning/deep learning, which highlight how these metrics varied in previous elections and how much weight they should carry in future elections.

The principal motivation for this project comes from the expansion in the data being recorded within the political system. Additional motivation comes from the potential of providing an alternative to traditional polling data for the purpose of forecasting election outcomes. In 2016, the majority of polls predicted that Hillary

Clinton would win the election and would win almost all of the battleground states[1] [1]. Explanations for those polls being wrong included overhyping Clinton in the battleground states and the underestimation of undecided voters that would vote for Trump . However, this paper examines and experiments whether polling is the best way of forecasting an electoral outcome, or whether we can use other data to more accurately predict elections.

Other motivations arise from the spread of news through social media and other avenues on mobile devices. The role of the television and the role of social media have played a big role in elections during recent years. In 2018, television was the main source of news about presidential election campaigns for the majority of adults. Social media was the second main source of news[2]. Although classifying sentiment from social media platforms such as Twitter is already possible, this project utilizes learning algorithms to evaluate how social media sentiment can be used in understanding election patterns.

Furthermore, this paper not only aims to explore the metrics mentioned above, but also emphasizes how these variables can explain, model, predict elections, or explain other political phenomena. This paper also advances the idea that voters, or anyone reading any statement online should be aware of the credibility of the source presenting the content. Therefore, this research also focuses on how we can algorithmically

---

[1] Cohn, Nate. "A 2016 Review: Why Key State Polls Were Wrong About Trump." *The New York Times*, The New York Times, 31 May 2017, www.nytimes.com/2017/05/31/upshot/a-2016-review-why-key-state-polls-were-wrong-about-trump.html.

[2] Gordon, Kyle. "Topic: Social Media and Politics in the United States." *Www.statista.com*, 2019, www.statista.com/topics/3723/social-media-and-politics-in-the-united-states/.

represent credibility, and how we can apply these learning algorithms in a way that people could use widely. To summarize, the paper aims to present how credible social media is in describing the popularity of candidates/forecasting elections, how credible are certain indicators like economic status and outcomes of previous elections to forecast future elections, and how can information on the internet be categorized as fake news or real news to influence users behavior patterns.

2. Introduction

Before going into the details of the proposal of how Artificial Intelligence can be used to improve the experience of voters and to discuss how Deep Learning ideas can be used to improving electoral forecasts, we must be able to understand the background of the role of data in modern politics, and an overview of deep learning/machine principles that we utilize in politics and other fields/industries.

2.1. Deep Learning

The idea of machine learning arguably extends back 70 years, when the idea of the Neural Network was first developed with the creation of the first Multilayer Perceptron(MLP) at Cornell University. It isn't very important to know the specifics of the architecture of the MLP in comparison to other types of algorithms, but it is important to note that the creator of it, Frank Rosenblatt, called it "the first machine which is capable of having an original idea" [3]. Since then, many mathematicians and computer scientists have worked endlessly on research to develop the modern field of data science, machine learning and artificial intelligence.

For the purposes of this paper, we focus on how machine learning and deep learning are applied rather than the theory and mathematical proofs behind their

[3] Lefkowitz, Melanie. "Professor's Perceptron Paved the Way for AI – 60 Years Too Soon." *Cornell Chronicle*, 25 Sept. 2019, news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon.

algorithms. In a rudimentary explanation of these topics, the idea behind artificial intelligence is to learn a pattern, prediction or even learn an experience without being programmed explicitly. By this, a computer can learn either one of these things from just performing patterns based on certain computational conditions. The big focus of machine learning comes from being able to gain experience from data. The golden rule regarding this is that the more data an algorithm is able to train on, the better it will perform. However, this holds mostly true if the data is uniform, and doesn't have many outliers or misinputs.

Within machine learning, there are three main types of learning algorithms, but for the purpose of this paper, only one is to be used. The three types of learning algorithms are Supervised Learning, Unsupervised Learning and Reinforcement Learning. Supervised Learning is a learning algorithm where the data used to train the model is completely labeled, and the learning algorithm attempts to predict future events based on this labeled data. An example could be a learning algorithm to predict whether an input image either contains a dog or a cat. To train this learning algorithm, a dataset composed of images being labeled as either containing a dog or containing a cat would be fed into the network, making it a supervised model. Unsupervised Learning is when the data used to train the learning algorithm is neither classified or labeled. An example of an unsupervised learning model would be to imagine a scenario of a dog playing with one person, and then playing with another person without knowing the second person. The way this happens in unsupervised learning is that the algorithms start to make features from the former inputs, so in this example, the dog learned some of the features of the

original person that made it want to play with the second person before ever seeing the second person. The learning algorithm in this scenario may or may not have the right prediction output, but the algorithm uses its learned features into categories that it has formulated from the training data. Finally, Reinforcement Learning functions off the idea of how an arbitrary agent reacts with an environment. A reinforcement learning algorithm learns by sampling actions available to the agent, where the agent receives a reward based on the action it has taken. Therefore, if the agent takes the wrong action, it gets a harsh negative penalty, and if the agent takes the right action, then it receives a favorable reward. Overtime, the learning algorithm learns optimal performances by how it receives rewards in a sequence [4].

For the purpose of this paper, we use Supervised Learning, as all of the data that will be used to train the algorithms will be labeled.

2.2. Big Data in Politics

The evolution of how data has been being used in politics is quite astonishing. To understand how it is being used, we must understand how it can be used in the idea of polling data, how campaigns use it to get a sense of what voters want to see in their elected official, how campaigns create models to find the likelihood of voting patterns, etc. Therefore, the question isn't how campaigns use big data in politics but why they use big data in politics. The primary reason campaigns use political data is to be able to

[4] https://ai.stanford.edu/~nilsson/MLBOOK.pdf

predict a future event with confidence [5]. This can either be how well a slogan or campaign rhetoric resonates with focus groups or the likeliness of a candidate winning an election. At a basic level, candidates collect this data to maintain a record of people in an area, a record of people that are likely to vote for a candidate, and a record of people that are undecided in whom they want to vote for. Furthermore, candidates and campaigns collect this data in numerous ways. This ranges from anytime a person signs up for an email subscription of a campaign, receives a text message from a volunteer of a campaign, or even answers a question to focus groups at voting booths. The way all of this information is gathered has changed over time, especially with the rise of the internet [6]. Before the widespread use of what we call the modern internet, campaign information regarding voters was limited to physical information such as voting history/voting patterns and geographical information regarding a voter. Now, there are many more metrics a campaign can use to get the complete picture of a voter. With the rise of the internet, there has been a rise of analytics companies that do analysis of voters for campaigns, and they are able to get a lot more information than just geographic information and voting history. For example, a very popular metric for creating a story of a person can be to see a person's clickthrough pattern on social media such as Facebook or Twitter. A clickthrough pattern is a metric that shows what a user clicks on in their social media app. It paints more of a complete story of a person because not only does it show everything a user clicks on, it can also predict how likely a user would click on

[5] Nickerson, David. *Political Campaigns and Big Data*. 2013.
[6] Körner, Kevin. "Digital Politics AI, Big Data and the Future of Democracy." *Deutsche Bank Research*, 2019.

other content given the fact that the user has clicked on similar content in the past. We saw the first major use of these tools in politics during the 2008 US Presidential Election between Barack Obama(D) and John McCain(R). Political analysts characterized [RBE1] this election as once in a generation type of election because of how data was used in this election, compared to how it was used in previous elections. Another key feature for the rise of political data is its usage in social media. Unlike presidential elections starting from 2008, campaigns in the 20th century had to maintain information about voters without the help of social media. Therefore, these campaigns were limited to physical voter data, and tracking these voters and their patterns often became inefficient. It is noted by many political experts that Barack Obama's 2008 campaign was the first political campaign to capitalize on the integration of social media in a presidential campaign. In fact, the Pew Journalism Center took note regarding the differences in social media between Barack Obama and John Mccain in the following quote: "Even after the McCain enhancements, Obama has more MySpace friends by a nearly 6-to-1 margin, more Facebook supporters by more than a 5-to-1 margin, twice as many videos posted to his official YouTube channel, and has more YouTube channel subscribers, by an 11-to-1 margin." This means that campaigns have evolved to become social media centric, so the belief is that social media can also describe the elections and possibly predict which candidate will win.

3. Related Work

With the tremendous rise of big data over the past two decades, data scientists and researchers have not only expanded the educational training regarding this field, but have also worked on projects dedicated to tackling problems in politics and to tackling shortcomings of current practices in forecasting elections. Examples of advancements include to the changes in how international trade is conducted, how lobbying is conducted, or even behavior patterns of certain voters. These behavior patterns tell a lot about a voter including his/her interests and how likely the voter is to click on certain links.

However, even with these advancements, the rise of big data has also led to a rise in some notable issues that are worth noting, including the rise in misinformation, and the possibility of algorithmic bias in search results and feed results. In this study, we dive into these two issues and propose a solution for the rise of misinformation and fake news.

3.1. Academia

During the 21st century, many educational institutions have been developing courses that not only focus on machine learning and artificial intelligence, but also on the applications of these two subjects. These applications include computer vision, natural language processing, robotics, etc. However, another application that is most appropriate

for the topic of this paper is the idea of data science in politics[7]. For example, the Massachusetts Institute of Technology (MIT) released an undergraduate course for the Spring 2020 semester called "Machine Learning and Data Science in Politics." There are other iterations of this course that are presented in other universities but this course gives insight to how politics is incorporated with machine learning to be taught at universities. The course begins by going into big data topics such as text processing and initial data analysis. Then, the course dives into learning algorithms, specifically in unsupervised learning, that focuses on clustering and networks such as an artificial neural network. Finally, the course goes over applications of data science and machine learning through textual analysis to show how these tools can be used in issues such as political voter info and user pattern tracking on social media. With the rise of social media and big data in the 21st century, universities have expanded knowledge in this field, furthering the commitment to solving more issues in politics through artificial intelligence.

3.2. Social Applications

Artificial Intelligence has come a long way since its initial research in the 20th century. Thanks to ideas such as cloud computing, social media, the internet and computing power, artificial intelligence has made its way to many areas of our lives[8]. Even though data science is becoming more integrated into society, the number of controversies that are growing with it are hard to ignore. The reason this must be said is

[7] http://www.mit.edu/~gtoral/machinelearning_syllabus.pdf
[8] Schatsky et al., 2014; Kelnar, 2016; Schwab, 2016; Huber, 2017; Makridakis, 2017

because with all of the advancements in technologies such as self-driving cars or robotics, people must also understand that it has been critiqued heavily in recent years over how transparent the understanding behind artificial intelligence has been in the eye of the public. This is particularly the case in Europe, starting in 2018 when the European Union created regulations aimed at restricting "automated individual decision-making." More specifically, the EU reasons that citizens have a "right to explanation" of how and why automated decision-making algorithms make certain decisions about people. This largely started a debate on the ethics of using Neural Networks for decision making, as Neural Networks have a difficult time showing the step by step process leading to a certain decision. Furthermore, these regulations are also a direct attack on large tech companies like Facebook and Google, which often use individual data for their own application features and advertising campaigns[9]. This, among many other conflicts, is how societies have combated the rise of artificial intelligence. Although the ethics and involvement of data science in society won't be covered in the scope of this project, it is still important to note that understanding how these algorithms work and the ethical applications behind these algorithms is extremely important.

Furthermore, artificial intelligence has been widely used by large tech companies, major news outlets, and government agencies at a massive rate over the past two decades. One major application of artificial intelligence is fake image detection on social media. Although this project does not relate to fake image detection , it is an important

[9] Metz, Cade. "Artificial Intelligence Is Setting Up the Internet for a Huge Clash With Europe." *Wired*, Conde Nast, 17 Oct. 2017, www.wired.com/2016/07/artificial-intelligence-setting-internet-huge-clash-europe/.

application of how artificial intelligence is applied to politics, and is one of many applications of AI in politics. Fake image detection takes an image as an input and outputs a probability value of how likely the given image is not real. The motivation arises from the spread of fake news on social media, especially in the form of images. Moreover, researchers were able to achieve about a 97% accuracy when testing their model[10]. Developments in computer vision and convolutional networks specifically have increased the amount of photo analysis. For example, current researches at Adobe have been working on how well computers can detect which pixels in an image have been altered and predict what the original image is[11]. Technologies like these have helped transform the impact artificial intelligence has on social media and society.

[10] Alshariah, Njood Mohammed, and Abdul Khader. "Detecting Fake Images on Social Media Using Machine Learning." *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, 2019, doi:10.14569/ijacsa.2019.0101224.

[11] Adobe Communications Team. "Adobe Research and UC Berkeley: Detecting Facial Manipulations in Adobe Photoshop." *Adobe Blog*, Adobe, 18 June 2019, theblog.adobe.com/adobe-research-and-uc-berkeley-detecting-facial-manipulations-in-adobe-photoshop/.

4. Improving Electoral Predictions with Learning Algorithms

In this study, we seek to algorithmically create methods able to accurately predict candidate popularity and election outcomes themselves. We first examine how social media, specifically sentiment expressed on social media, can be used to forecast a candidate's approval rating or forecast an election itself. Second, we will utilize feature selection and feature engineering to predict election outcomes through learning algorithms. This study focuses on more of the use of statistics and indicators to predict an election than on the use of social media sentiment.

4.1. Using Sentiment Analysis to Model Polling Patterns

Social media has become a powerful tool in not only getting information on potential voters, but also to influence them. The purpose of sentimental analysis over the Twitter platform is to see how sentiment over different topics is modeled in different states across the country. Although sentiment analysis can be performed over a multitude of topics, the topics have been narrowed down around the queries of "Biden" and "Trump." This focus will allow us to see an analysis of sentiment on a national level composed of all of the states, and on a swing-state level. A swing state is a state in the United States where the Democratic and Republican party have a similar popularity, making the state very competitive to win during an election. Compared to states like California and Texas, where California usually is won by the Democratic party and Texas

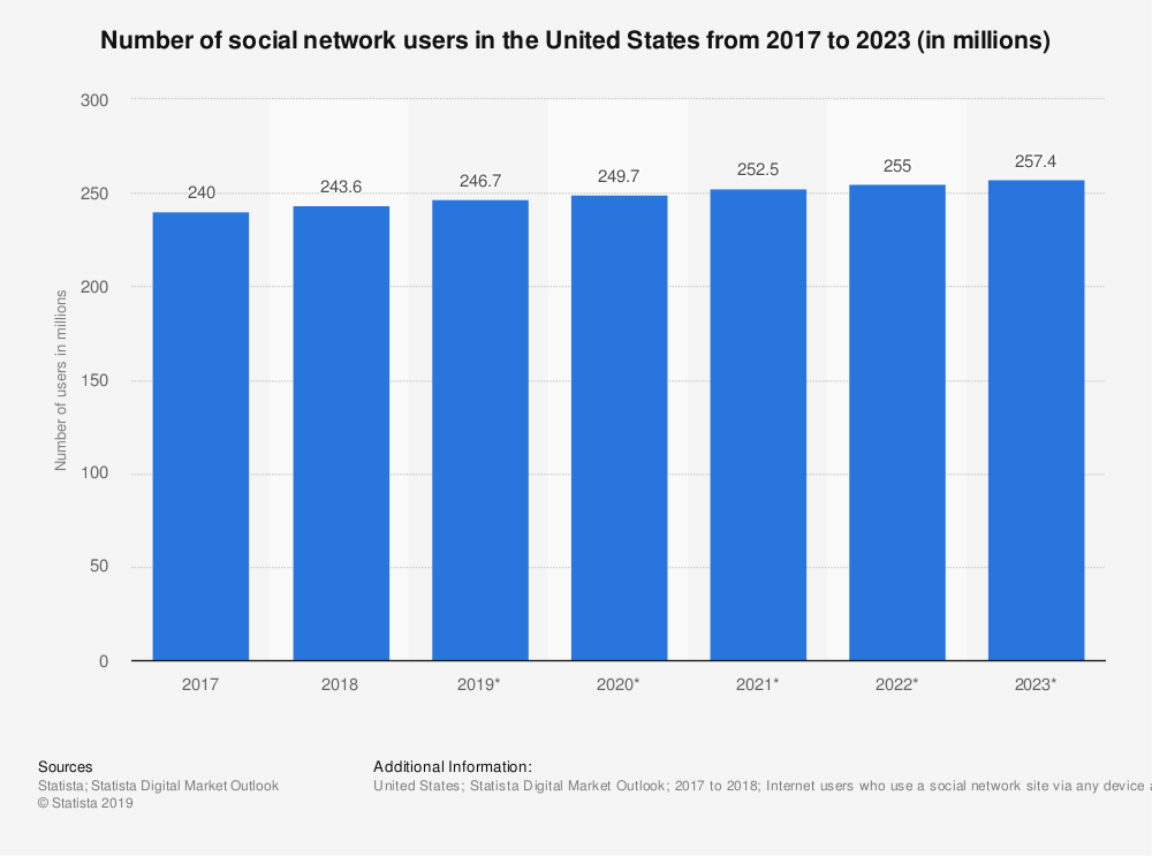
is usually won by the Republican party, swing states are toss ups and are usually won by a small margin[12]. Then, we will see what the sentiment analysis looks like in comparison to current polling data provided by RealClearPolitics.

Sentimental Analysis is usually used to convey the sentiment of some textual document or idea. However, it can also be used to predict mood from image-based data. Using machine learning models, we can train models using data about sentiment from prelabeled sentiment and can present it by analyzing the mood of articles or posts on social media.. In this project, tweets are either classified as positive or negative regarding a candidate.

To do this, data is collected from Twitter regarding tweets that mention the term "Trump" or "Biden". Tweets are then preprocessed to remove words or parts of speech that could be interfering with model training. For example, the words "the" and "a" are very commonly used. Therefore, these words are filtered out to not add noise to the data. After this, the model is trained with multiple learning algorithms. After the training, the model is evaluated by testing data and is presented in visualizations that show sentiment intensity on a map. By this, the map will be colored in correlation to the scale of polarity predicted from the sentiment classifier. The results of the chosen model are then compared to the current polling of a certain state. The results will also be compared to the candidates' national polling percentage as well. The fact is that more and more Americans are using social media and that Americans are getting their news from social media. The Pew Research Center published in late 2019 regarding the percentage of adults that get their news from social media. The study concluded that the percentage of

[12] https://www.merriam-webster.com/dictionary/swing%20state

adults who frequently got their news from social media increased from about 28% to 55% [13]. The increase in the number of social media users in the US from year to year also helps explain this trend. The number of users in the US is projected to be about 257.4 million by 2023[14]. The hope is that the analysis of social media can illustrate the opinion of voters in swing states and in the nation.



Furthermore, the motivation for this subject arises from the following questions. How do sentiments from states correlate with the popularity or polling of certain candidates? Are

[13] Suciu, Peter. "More Americans Are Getting Their News From Social Media." *Forbes*, Forbes Magazine, 11 Oct. 2019, www.forbes.com/sites/petersuciu/2019/10/11/more-americans-are-getting-their-news-from-social-media/#3345507b3e17.

[14] https://www.statista.com/statistics/278409/number-of-social-network-users-in-the-united-states/

there any relations between the sentiment surrounding certain candidates on social media?

4.1.1. Data Collection

The model that is used for the sentiment classification comes from a popular hub for data scientists to find datasets. The name of the dataset is Sentiment140, and it is a collection of 1.6 million tweets with a classification of sentiment for each of the tweets. This dataset is publicly available online. All 1.6 million tweets are used in training the model[15]. Other datasets were also evaluated, but this dataset was eventually chosen as the data was well balanced in terms of the positively and negatively labeled tweets it had. Out of the 1.6 million tweets, about 70% of the entries were used for training of the model and the other 30% of the entries were used for testing/validating the model.

Data was then collected to visualize the comparisons between Donald Trump and Joe Biden in swing states and in the nation. This data was collected from Twitter feeds by using the Twitter API.

4.1.2. Issues With the Twitter API

The goal of using the Twitter API was to represent the sentiment from all of the tweets collected in a geographical map; however, there is no guarantee that people who tweet have a specified geotag location on their tweets. To solve this, the location

[15] https://www.kaggle.com/kazanova/sentiment140

parameter was used to only query the tweets that had a location specified on the tweeter's account. Another concern was that some users created fake locations, or locations that could not be traced to a geographical location. For example, some twitter users had their locations set as "America, USA", "🌳🇺🇸Tree of Liberty, USA 🌳🇺🇸", "'merica", or even "Narnia." Since these tweets could not be used in creating an accurate representation of the population, these tweets were disregarded from the ones collected by the api. The issue with this method is that only so many tweets can be collected before the credit limit is reached to retrieve tweets. Unfortunately, the request for a premium Twitter developer account was rejected, so tweets were gathered in batches of roughly 2,500 per request. The other issue was that the standard Twitter API only allowed for a certain number of requests before the server bans requests for an amount of time. This made it more troublesome as only 10,000 tweets were collected in 30 minutes. Another big issue was that each request to get tweets sometimes returned some overlapping tweets, therefore the process of storing tweets programmatically filtered only unique

4.1.3. Preprocessing

Since the dataset is composed of tweets, each tweet needs to be filtered of objects that would affect its training. Tweets compose of the username that is tied to whoever creates a tweet, the actual content of the tweet, urls that could be attached to a tweet, and hashtags a user chooses to add. For any given tweet, we want to preprocess it so that it looks like an ordinary sentence. Therefore, when we preprocess a tweet, all links are

removed, the username of the tweeter is changed to “AT_USER,” and all symbols/special characters are removed.

The following is what a tweet looks like before any preprocessing:

RT @Lawrence: Trump tariffs are paid by American consumers not by foreign countries.

The following is what a tweet looks life after preprocessing:

AT_USER Trump tariffs paid American consumers foreign countries

4.1.4. Methodology

In this scope of the project, different learning algorithms are used to train and test the models. This is done to optimize model performance as different learning algorithms can yield different performances. A Logistic Regression(LR), Neural Network(NN) and a Long-Short Term Memory(LSTM) memory were the tested algorithms to create the model. After choosing the learning algorithm that performed the best

A Logistic Regression model is a supervised classification model where it models the inputted data using a sigmoid function rather than a linear function. The benefits of using a logistic regression is that it can separate non-linear data, but will perform suboptimal on linear data due to the activation function. Another benefit of using a logistic regression is that it is computationally inexpensive compared to other algorithms such as a neural network.

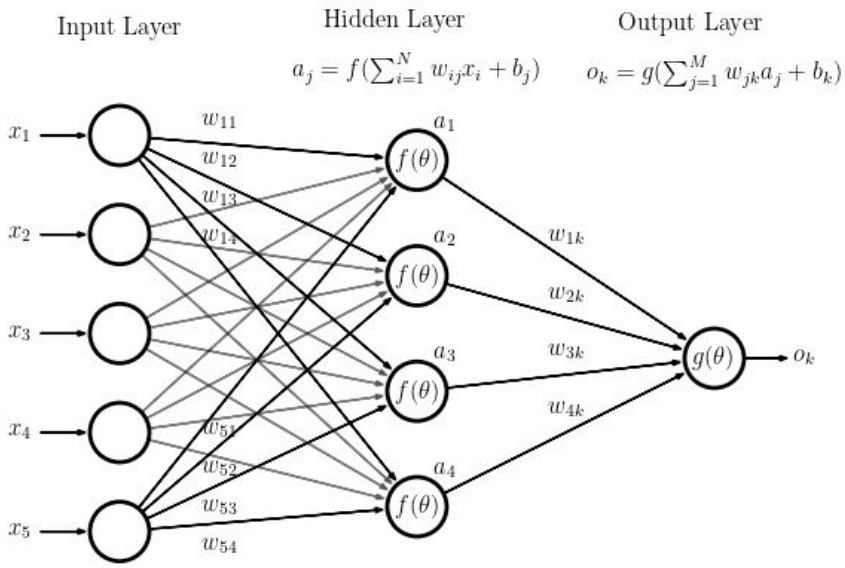


Diagram of Neural Network[16]

The Neural Network, which has been very popular as of late, has completely changed the impacts of Deep Learning in the 21st century. In summary, the purpose of a neural network is to computationally present the decision process of a human being. The image above shows how a Neural Network generally works. Basically, input gets transferred between the input layer to hidden layers. Depending on the setting for hidden layers in the network, they can vary from a single hidden layer to multiple hidden layers. The initial inputs are used to compute weights that are passed along through the hidden layers, where the neurons go through the activation function. Depending on the number of activation layers, the neurons either are forwarded to the output layer or go through more hidden layers.

[16] https://www.astroml.org/_images/fig_neural_network_1.png

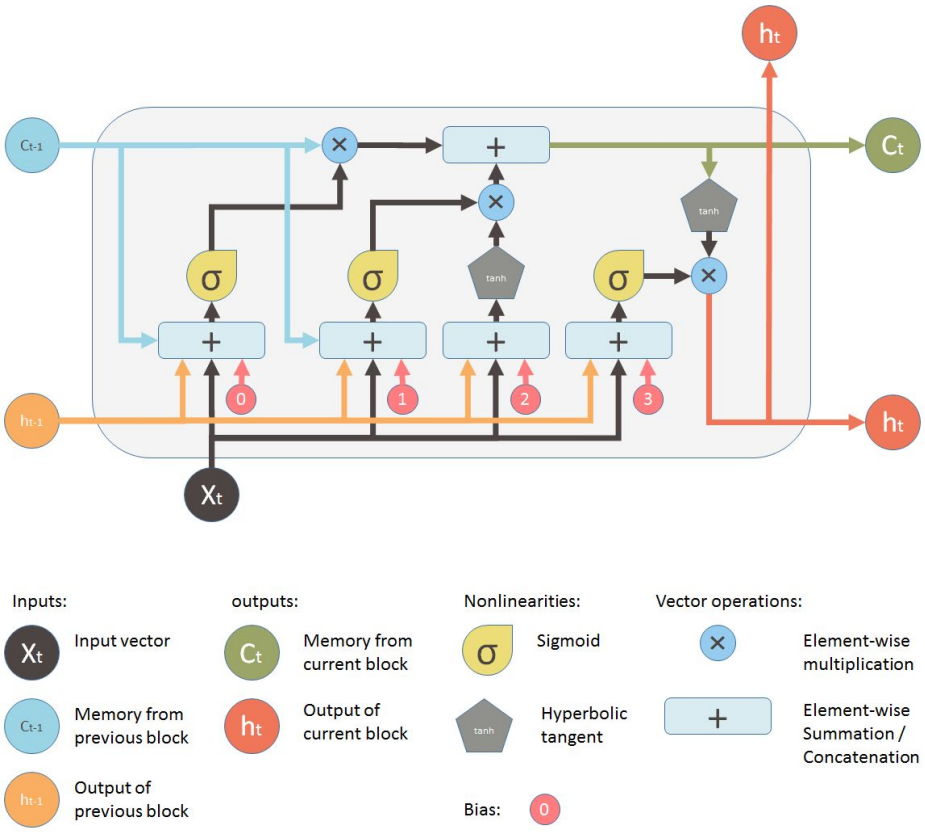


Diagram of LSTM[17]

An LSTM is a type of Recurrent Neural Network that can learn long term dependencies. Imagine the sentence, "Thomas, a 24-year old man lives in San Francisco. He has a female friend Lilly. Lilly works as an engineer for Facebook in New York whom he met recently in a school alumnus meet. Maria told him that she always had a passion for _____." In this sentence, a long-term dependency refers to the gap between the information wanted and the place it needs to get predicted. Moreover, an LSTM was used for the classification because LSTMs work well with speech recognition, sentiment analysis, language modeling, text prediction, etc. More

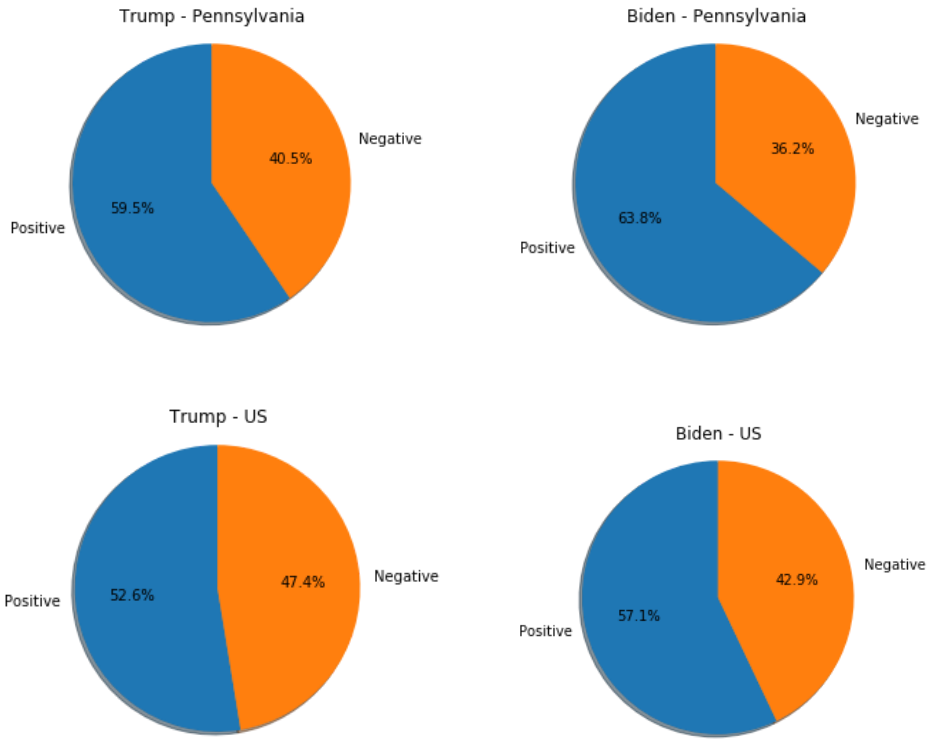
[17] https://devopedia.org/long-short-term-memory

information regarding the model is shared below. The following is the recorded hyperparameters configured for the LSTM model, which had the best performance.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 28, 128)           256000
_________________________________________________________________
spatial_dropout1d_1 (Spatial (None, 28, 128)           0
_________________________________________________________________
lstm_1 (LSTM)                (None, 196)               254800
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 394
```

4.1.5. Results

When testing all of the mentioned learning algorithms, the LSTM performed better than the Linear Regression or Neural Network. There were two runs of training the model. In the first run, only 512,000 tweets were used and the split between positive to negative tweets was 1 to 2 respectively. The accuracy for this run was about 82% with the accuracy for identifying negative sentiment being higher than the accuracy for identifying positive sentiment. This was no surprise as the training data split reflected the outcome of the model. Moreover, the initial run with the decrease in data was due to the limitations in computational power. However, when using the entire data set, the model had accuracy of 96% and it had optimal results with both positive and negative sentiment.



Results of tweets collected from a swing state such as Pennsylvania and tweets collected from any part of the United States are shown above. 10,000 tweets were used for each candidate in the Pennsylvania focus, and 20,000 tweets were used for each candidate in the national focus. This is shown to compare how Donald Trump and Joe Biden match at a swing state and national level. Those results are then compared to current polling from RealClearPolitics to examine how much credibility sentiment analysis holds for predicting correct polling numbers and correct electoral forecasts. For the state of Pennsylvania, RealClearPolitics reported its last poll done in Pennsylvania(Harper(R)) to have Biden leading by 6 points. From these results, we can see that Biden has an advantage of 4.3% in terms of positive sentiment. Nationally,

FiveThirtyEight reports that Biden has a 6.4% advantage, while the results of this research show that Biden has a 4.5% advantage in national positive sentiment.

4.1.6. Conclusions

From the results shown above, there appears to be a similarity between sentiment analysis and the polls. However there are a few things that must be acknowledged regarding the sentiment from Twitter. One, the average Twitter user population might not reflect the average voting population. This means that even though the measurements were somewhat similar, this methodology doesn't guarantee consistency as enough tweets were not streamed. Furthermore, the biggest limitation in the prediction of the election came from how many tweets were streamed from the Twitter API, and how many tweets were filtered out due to issues with location are visualized on the map because of the problems with tweet location. Due to the relatively small amount of data because of Twitter API request limits and the inability to use every tweet based on location restraints, there were discrepancies between the sentiment prediction and the polling data. Finally, it's possible that tweets about Biden or Trump doesn't necessarily characterize how people could feel about either candidate. A person can easily criticize either candidate on Twitter or other social media platforms, but vote for the other candidate. When tackling this problem in the future, having more stream data from Twitter could improve the results of the predictions; however, it could also add more randomness at the same time. This is because there are many details still unknown about the quality of tweets. For example, some of these tweets can come from bots or internet

trolls. Therefore, there's a chance that sentiment analysis won't always yield similar results to polling.

4.2. Using Feature Engineering to Forecast Elections

Researchers have been addressing the issue of predicting presidential elections through algorithmic means for many years now[18]. It is also known the forecasting elections is also a very difficult issue due to the limitations in the amount of data available for this. The other issue is the numerous features available that could feed into influencing election outcomes. For example, when we think about which factors can help a candidate win an election, there are many competing views. Some can argue it's a candidate's likeability or how the public perceive him/her as a person; others can argue that the state of the economy is a big factor in predicting elections, or even the amount of funding a campaign has[19]. Among these, people are also divided on what role does the United States Congress play on the outcome of presidential elections. Some believe that the outcome of the last midterms election can play a big role in an upcoming presidential election. The purpose of this research is to evaluate how different learning algorithms perform when forecasting elections. The other purpose is to evaluate which features are most important when predicting elections.

[18] Abramowitz, et al. "Modeling and Forecasting US Presidential Election Using Learning Algorithms." *Journal of Industrial Engineering International*, Springer Berlin Heidelberg, 1 Jan. 1988, link.springer.com/article/10.1007/s40092-017-0238-2#Bib1.
[19] Fair R (2011) Predicting presidential elections and other things. Stanford University Press, Stanford

4.2.1. Literary Analysis of Current Methods

Forecasting elections goes back to the 1970s when economists and social scientists like Roy C Fair forecasted elections by simply observing the status of the economy[20]. Other researchers followed in forecasting elections by other metrics such as using the outcomes of the previous elections to predict the current election, using the president's job approval rating, Some other researchers such as Hollbrook and DeSart utilized the percentage of voters that voted for a specific party in the last election[21]. Furthermore, in this attempt of forecasting elections, we will see how a combination of these features with additional features will influence presidential predictions. We will also see how deep learning methods improve prediction outcomes. The main point to understand from this is that election prediction is not a novel idea, and that researchers have been proposing solutions for at least the past 30 years. Therefore, this research proposed to evaluate how different machine learning and deep learning expand on previous methods, and how feature engineering can lead to an improvement in accuracy.

4.2.2. Data

Unfortunately, there is no centralized dataset that contained all of the features that were used in training the model; therefore, data was collected from a variety of sources to construct the dataset. Many of the quantifiable data that was used in this dataset came from research done from Gallup, RealClearPolitics, etc. To illustrate the data that was

[20] Fair RC (1978) The effect of economic events on votes for president. Rev Econ Stat 60:159–173
[21] Holbrook TM, DeSart JA (1999) Using state polls to forecast presidential election outcomes in the American states. Int J Forecast 15:137–142

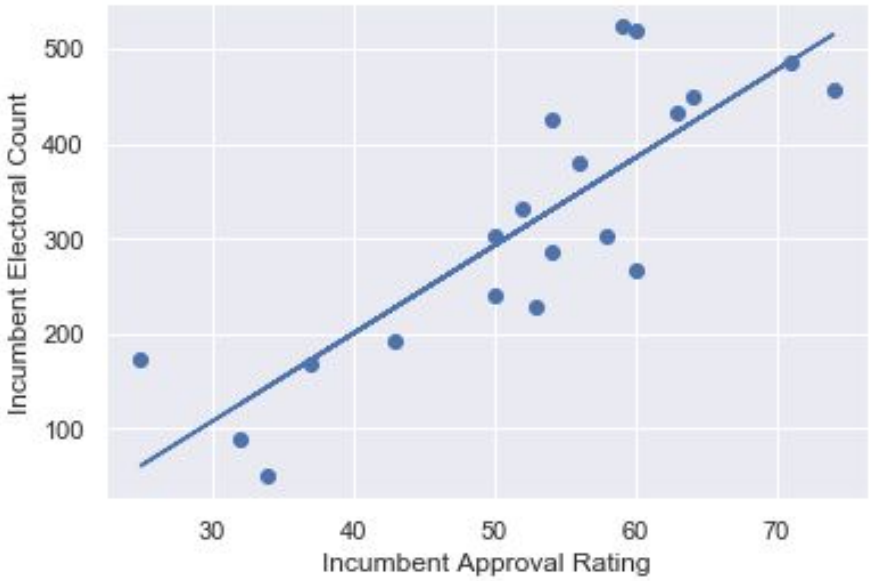
used in training the model, the following columns were created: the incumbent party, the number of years the incumbent party has been in office, whether or not the incumbent party won the last house/senate election, how many seats did the incumbent party hold in the house/senate after the last election, the presidential approval rating of the incumbent president during October of the election year, the unemployment rate, the gdp growth rate, inflation rate, whether the country is facing an economic recession and whether the country is in a national crises. These variables are then used to predict whether or not the incumbent party has won the election. The reasoning behind choosing these variables is the following. There are a few variables that have strong ties with the outcome of the election. For example, if the economy is doing very poorly under the incumbent party, the expected outcome is that the public would want to elect a candidate from the opposing party. This can be seen from Barack Obama's popularity during his 2008 campaign compared to his opponent, John McCain. Another example is that the approval rating and support for a sitting president increases in the time of a national crisis. For example, George W Bush's approval rating was roughly around 90% after the tragic event on September 11, 2001. His popularity was relatively popular during the 2004 election as the nation was at war with Iraq. Furthermore this dataset contains every presidential election from 1940-2016. This dataset starts in 1940 because that is a good starting point for modern American politics. In 1940, President Franklin Delano Roosevelt was seeking reelection in the midst of The Great Depression and World War II. Another reason for this year is that the early twentieth century experienced a sharp rise

in globalization as countries became more dependent on each other. Furthermore, the progressive movement also expanded during this time in the United States.

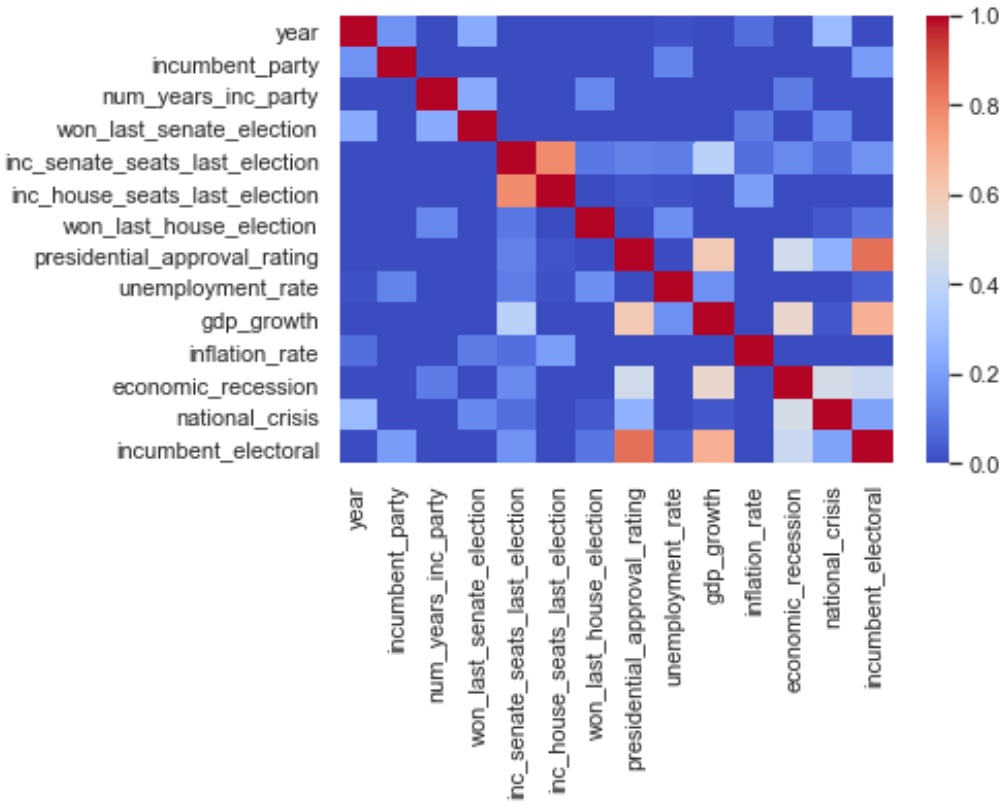
There are 21 total data points in this dataset, from which 17 were used to train the model and 4 were used to test the model. Unfortunately, the data is very limited due to the number of elections that can actually be used to train/test the models. For example, elections before 1900 would not be ideal as the issues that surrounded politics back then are different from the issues that are prevalent today.

4.2.3. Data Analysis

When creating the dataset, there were a few notable things to understand when thinking about what influences the outcome of an election. For example, the Democratic party had a majority in the US House of Representatives from 1953 - 1994, but the party that controlled the white house changed numerous times between those years. This means that the results of the previous US House elections had little contribution in influencing the outcome of a presidential election. Another interesting but logical fact about the dataset is how closely the incumbent presidential approval ratings and the incumbent electoral count for an election are. In fact, presidential approval ratings are the most influential feature on the outcome of elections. A visual of this phenomenon is provided in the graph below.



Furthermore, a correlation of how the features chosen in this dataset are shown below.



4.2.4. Methodology

In this scope of the project, different learning algorithms are used to train and test the models. This is done to optimize model performance as different learning algorithms can yield different performances. Decision Trees, K-Nearest Neighbors, and Neural Networks were the learning algorithms to create the model. A Neural Network was described in a previous section of this paper, and a brief overview of the algorithms will be provided below.

A Decision Tree is a machine learning model that essentially creates questions to split what is an optimal value and what is not. Furthermore, a decision tree makes decisions without making any underlying assumptions about it. For example, a question for this dataset can be "is the incumbent president's approval rating greater than 60?" If the value of presidential approval rating is greater than 60 for a specific data point, then it will go in a different branch than if the presidential approval rating was less than 60. Since this research focuses more on the application of machine learning rather than the intuition behind it, it is fine to just understand these algorithms in basic terms.

The K-Nearest Neighbors is another machine learning algorithm that is used to train the model. The basic idea behind K-Nearest Neighbors is to find the relative difference of an arbitrary data point with all of the other data points in the dataset. Once that is done, the algorithm takes the top k data points and uses the majority output of those k data points. Once it takes the majority, it uses it to predict the output of the given data. For this research, k is set to three; therefore, the three closest elections are used to predict who wins a given election.

4.2.5. Results

The three learning algorithms were used to train the model. To test the model, the testing set consists of the 2004, 2008, 2012 and 2016 election. To provide a brief summary in how the models performed, the decision tree performed suboptimal as it only predicted one of the four elections correctly. Next, when the K-Nearest Neighbors(KNN) algorithm was used, it predicted one of the four elections correctly when k was set to 2, but it predicted three out of four elections correctly when k was set to 3. When training the model with the Neural Network, it also got three out of four elections right. However, the difference between the KNN model and the NN model is that when evaluating the complete testing accuracy, the NN has a testing accuracy of 70.8% while the KNN has a testing accuracy of 54%. The complete testing accuracy is evaluated, when the validation iterates over three random data points in the dataset to test the algorithm. Therefore, from testing, the Neural Network had the optimal performance. When trimming down the dataset to only features that had a correlation of at least 0.45, the accuracy was still relatively the same for all of the learning algorithms with the Neural Network performing the best.

4.2.6. Conclusions

Although this research is just one example of many trying to show how possible election prediction through machine learning is, it is important to understand that even if a model can achieve 80% plus accuracy, it is not possible to say how reliable it is due to

the restriction of data available for this topic. Furthermore, the results from this research show that election prediction through machine learning and deep learning methods is possible; however, it is still limited due to the amount of data available. This does not mean that the methodology in this research or other papers is bad, but that results can be promising, but can't guarantee a small variation in results. It can also be concluded that presidential approval rating is the most influencing variable while the outcomes of the most recent US House of Representatives variable is one of the least influencing variables.

5. Misinformation and the Spread of Fake News

Social media has never been as popular as it is today. It can also be argued that social media is the largest outlet for sharing news. However, this comes at a big tradeoff. Even though social media is low cost and very efficient to receive news, it is open for anybody to post on. This means that many people can post things on social media that might or might not be credible. Social media has become the largest source of fake news. The other issue at hand with this is that there isn't a widespread blocker or checker that eliminates this notion of "fake news." This project examines fake news detection on traditional news sources. This includes presidential statements, town hall meetings, etc. This project doesn't specifically target fake news on social media because detecting fake news on social media becomes even more difficult as detection algorithms must deal with information including profile engagements. In this research, we will experiment with fake news detection through machine learning; the liar dataset is used to train and test the models. The purpose of this section is to provide an algorithmic model that can be used as an extension to a browser. For example, when a user comes across a given article, a browser extension can provide a credibility rating (ex: 72% true) that can help a user avoid information with very low credibility and evidence.

5.1. Introduction

"Fake News" is a very hot topic currently. The idea is to computationally create a model to predict what is fake news and what is not fake news. The goal of this project is

to provide a stepping stone for an idea that could be used in a mobile/web setting when a user browses the internet and wants to know about how trustworthy a source is. This project aims to use Natural Language Processing to detect fake news, based on the content and other metadata found in the news articles. Although this isn't a new problem, it is a problem that can be addressed through algorithmic means. Fake news plays a big role in the information people receive, especially the information received on social media. For example, during the 2016 election, it was reported that Russia created fake accounts and social bots to spread false information on platforms such as Facebook and Twitter. In this section, I will overview fake news detection, present an algorithmic solution to detecting fake news, and address some individual research questions. "Fake news" is defined as an article, resource, etc. providing false information holistically. This does not mean that any article with a piece of false information is considered fake news. In this section, it is not physically determined what is fake and what is not fake, but rather the LIAR dataset provides a benchmark to what is fake so the models can predict if any given test is fake news or not. When thinking about this topic, there are a few things one should think about. There is no real definition of fake news. How does fake news differ among different people? For example, some fake news articles won't be regarded as fake to some people, but other people will mark them as fake news. With this being said, how can people create a definition that is agreed upon by the general population, so we can create better classification methods to detect fake news? How do the accuracies of these models improve when adding various inputs for training? For example, how do accuracies change when adding inputs such as the speaker, political party, etc? Is the

statement of the speaker even important when determining a given text as fake news or not? By this, is the text itself important, or is it the statistics of the speaker and context that makes the given text fake or not? Going along with this, do the attributes of the data (context, party, job) even significant, or is the history of the speaker the only reliable attribute? Lastly, how can we use these models to discourage reporting of fake news?

5.2. Data

The LIAR dataset is used to train and test the models[22]. This dataset is composed of 12.8K manually labeled short statements from POLITIFACT.COM. This dataset contains 13 columns, including the label of the statement, the statement, the affiliated political party, the speaker, etc. When examining the dataset, I noticed that it is well balanced in terms of the political party of the speaker and output labels of the statement. During the phase of this project, I used the label of the dataset a bit differently from the original intentions of the dataset. Originally, the dataset had multiple categories for its labels: true, mostly-true, half-true, barely-true, false, and pants on fire. When I first examined the dataset, I intuitively recognized that there could be a large overlap of statements between the different labels, especially between the labels of barely-true and false. I believed this caused the accuracies in the original paper of the dataset to be around 27%. Due to this belief, I split the labels into two categories during the pre-processing: true and false. True, mostly-true, and half-true were put into the true

[22] Wang, William Yang. "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection." *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, doi:10.18653/v1/p17-2067.

category, and barely-true, false and pants on fire were put into the false category. I did this to evenly split the categories among the two. When using the dataset in the models, I dropped the columns that pertain to the 6 true/false categories because I categorized it into two. Instead, I use an input variable called “Credit History” that is a computational metric provided by the LIAR dataset. “Credit History” is the numerical value for how trustworthy the speaker is by getting the counts for inaccurate and accurate statements. I believe this input variable will be the best predictor for whether a statement is fake or not. More information about the dataset will be noted in the Exploratory Data Analysis section of the paper.

5.3. Methodology

5.3.1. Preprocessing

Data is preprocessed by removing stop words, punctuation, etc to make sure all of the data is consistent. This is similar to the way tweets are preprocessed in the sentiment analysis section. The outputs of the data were also categorized from the six original categories into two categories: true and false. The other changes in the data include replacing the individual counts of the six labels (true, barely-true, etc) with the credit history.

5.3.2. Data Analysis

Before conducting the models, data analysis was also conducted to get more information that could be used in predicting whether given inputs are true or false. The following is the information found in the dataset.

Top 10 Subjects: Healthcare, Immigration, Elections, Education, Candidates Biography, Economy, Guns, Jobs, Federal Budget, Energy

Top 10 Speakers: Barack Obama, Donald Trump, Hillary Clinton, Mitt Romney, Scott Walker, John McCain, Rick Perry, Marco Rubio, Rick Scott, Ted Cruz

Total Number of Democratic Posts: 4,150 Total Number of Republican Posts: 5,687

Top 5 Truthful Speakers (True/False Ratio): Bill Clinton(0.81), Rob Portman(0.79), Tim Kaine(0.77), John Kasich(0.74), Hillary Clinton(0.74)

Worst 5 Truthful Speakers (True/False Ratio): Democratic Congressional Campaign Committee(0.12), Ben Carson(0.16), Rush Limbaugh(0.17), Michele Bachman(0.22), Donald Trump(0.26)

5.3.3. Features

Models were run to evaluate the effect of different inputs. The belief was that "credit history" was the most significant input. To test this, models ran with just credit history, the statement without anything else, the metadata (job, speaker, etc) and metadata, and then the credit history with statement and metadata.

5.4. Modeling

Above is the structure for the models. The baseline is that the model predicts every output by the majority of the true/false labels from the dataset. This simply implies all statements are fake. The Logistic Regression used a one hot encoder to process the text. From that, the Logistic Regression for the text was then used for the Logistic Regression model(Metadata), the SVM(Metadata) and the Decision Tree(Metadata). I also created a CNN and Bidirectional LSTM that ran using only the metadata without anything else, and the text without anything else. I did this to test how reliable the statement alone can be, and how reliable the metadata without the credit history can be. In the deep learning models, I created vectors for statements from a vocab dictionary, word embeddings with Glove embeddings layer. For the metadata in the deep learning models, I took the k most frequent items in different metadata columns and used them to create ids for the inputs, which I trained in the models.

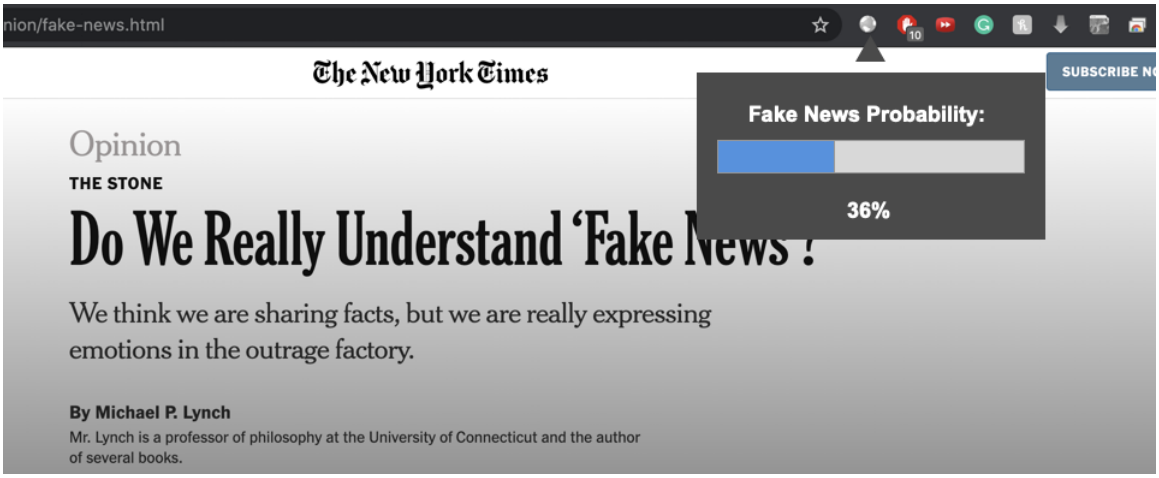
5.5. Results

From this liar dataset, the most important feature is credit history, which is about the best it can do compared to the whole metadata. Compared to the baseline, all of the other models performed better. The Logistic Regression, CNN and Bi-LSTM performed relatively close to each other. However, when I stacked the models and focused on credit history, the models improved in terms of performance (73.95% accuracy on the test set). The credit history is not so powerful in the training set, but it's more powerful in the validation set and over-powerful in the testing set. The reason is that in this data set, it aggregates the credit history from time to time, and the original way to split training, validation, and testing was according to the temporal order. Therefore, intuitively, as we know a person better, and we have a better idea to determine if he/she is inclined to lie or not. This is exactly the thing that happens here: as we accumulate more and more data, we have more power to determine if this person is going to lie.

5.6. Conclusions

From the model, we can conclude that the credit history of a person is a very good indicator compared to the actual statement from the person and compared to the metadata for the context of that statement. Furthermore, the purpose of this research is to promote the use of fake news detection algorithms when browsing articles on the internet. Furthermore, the use of these algorithms can help discourage the spread of fake news in numerous ways. One example of this can be a plugin built into a browser like Chrome. A

mockup design is provided below of how detection models can be implemented to encourage people who browse the internet to avoid fake news articles. The plugin would essentially present a probability of how likely an arbitrary article can be conveying fake news. This would allow users to receive information on every article that they visit. Moreover, this is one of many examples of how fake news detection algorithms can be used to discourage the spread of fake news.



6. Final Thoughts

From these three research topics, we can see how important the impact of machine learning and artificial intelligence has in politics. Ranging from summarizing public opinion regarding candidates through social media to forecasting presidential elections using metadata, we can see that the impact it has made will only continue to grow. Throughout these studies, one important current shortcoming, which is also apparent in many research projects, is the amount of data publicly available. This led to promising results in forecasting presidential elections, but also couldn't guarantee the consistency of the model. This issue also led to issues with sentiment analysis as there's a lot of variance when there is not enough data. Furthermore, this research will be expanded in the times to come. Perhaps not in the same scope as these projects, but probably ideas very similar.

Abramowitz, et al. “Modeling and Forecasting US Presidential Election Using Learning Algorithms.” *Journal of Industrial Engineering International*, Springer Berlin Heidelberg, 1 Jan. 1988, link.springer.com/article/10.1007/s40092-017-0238-2#Bib1.

Adobe Communications Team. “Adobe Research and UC Berkeley: Detecting Facial Manipulations in Adobe Photoshop.” *Adobe Blog*, Adobe, 18 June 2019, theblog.adobe.com/adobe-research-and-uc-berkeley-detecting-facial-manipulations-in-adobe-photoshop/.

Alshariah, Njood Mohammed, and Abdul Khader. “Detecting Fake Images on Social Media Using Machine Learning.” *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, 2019, doi:10.14569/ijacsa.2019.0101224.

Cohn, Nate. “A 2016 Review: Why Key State Polls Were Wrong About Trump.” *The New York Times*, The New York Times, 31 May 2017, www.nytimes.com/2017/05/31/upshot/a-2016-review-why-key-state-polls-were-wrong-about-trump.html.

Gordon, Kyle. “Topic: Social Media and Politics in the United States.” *Www.statista.com*, 2019, www.statista.com/topics/3723/social-media-and-politics-in-the-united-states/.

Körner, Kevin. “Digital Politics AI, Big Data and the Future of Democracy.” *Deutsche Bank Research*, 2019.

Lefkowitz, Melanie. “Professor's Perceptron Paved the Way for AI – 60 Years Too Soon.” *Cornell Chronicle*, 25 Sept. 2019, news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon.

Metz, Cade. “Artificial Intelligence Is Setting Up the Internet for a Huge Clash With Europe.” *Wired*, Conde Nast, 17 Oct. 2017, www.wired.com/2016/07/artificial-intelligence-setting-internet-huge-clash-europe/.

Nickerson, David. *Political Campaigns and Big Data*. 2013.

Schatsky et al., 2014; Kelnar, 2016; Schwab, 2016; Huber, 2017; Makridakis, 2017

Suciu, Peter. “More Americans Are Getting Their News From Social Media.” *Forbes*, Forbes Magazine, 11 Oct. 2019, www.forbes.com/sites/petersuciu/2019/10/11/more-americans-are-getting-their-news-from-social-media/#3345507b3e17.

Wang, William Yang. “‘Liar, Liar Pants on Fire’: A New Benchmark Dataset for Fake News Detection.” *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, doi:10.18653/v1/p17-2067.