Shaquiel Pashtunyar
DSC680 Project 2
Milestone 2

Shaquiel Pashtunyar
DSC680 Project 1
Milestone 2
Draft Paper

During the 2016 presidential election and all through the following years we saw a rise in fake news. One article from Yale stated, “There are hundreds of fake news websites out there, from those which deliberately imitate real life newspapers, to government propaganda sites, and even those which tread the line between satire and plain misinformation.” A big reason for the rise in fake news is that there is a low barrier of entry to get the disinformation out there. This is true with the rise of social media. We can see in one chart for Visual Capital that while social media plays a role in 10% of all news traffic, it represents 42% of all fake news.

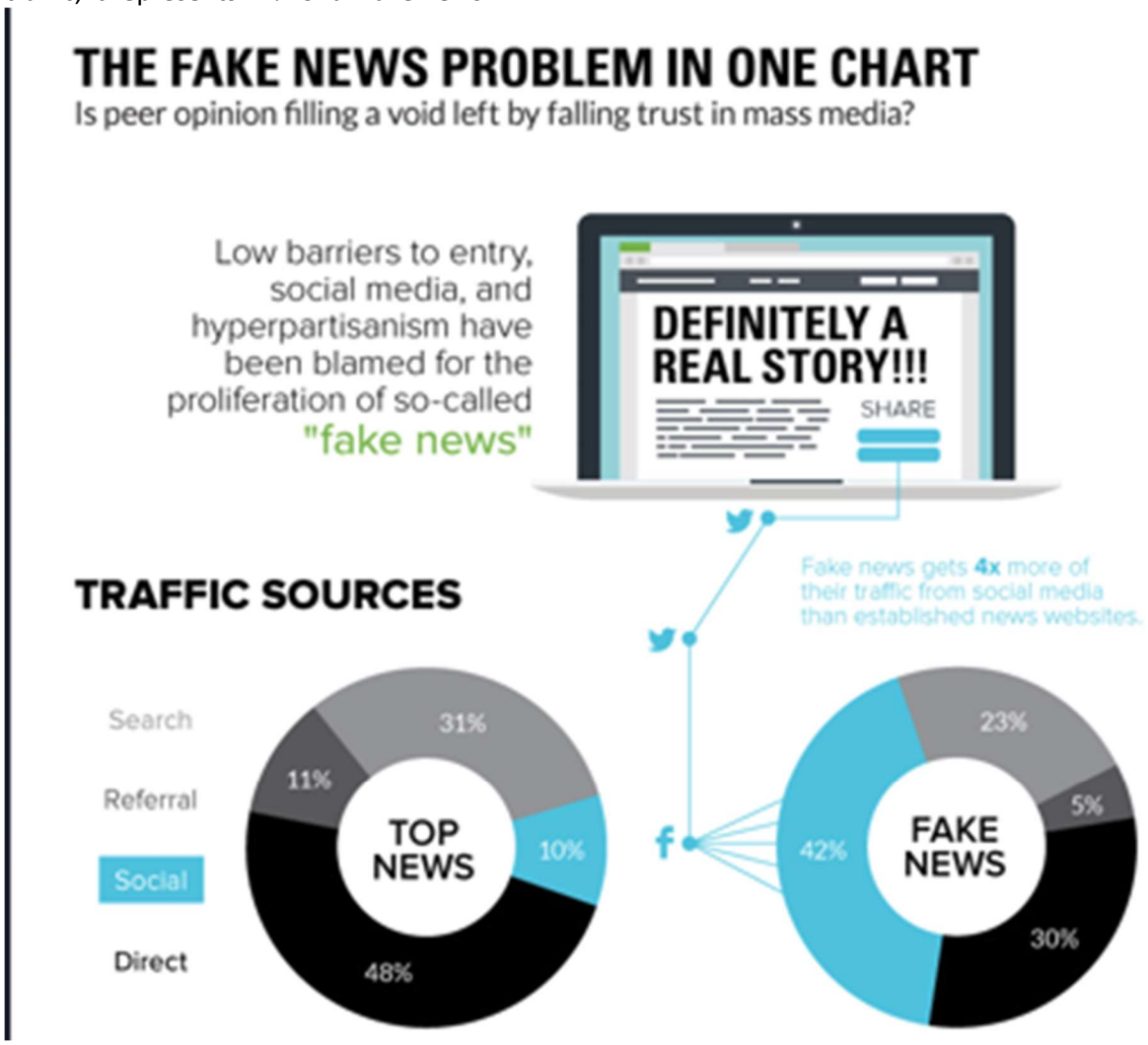

Figure 1 We see that social media makes up most of the fake news out there.

One more statistic we see is that with the rise in social media, peoples trust in social media is at an all-time low. This is true when both sides of a political spectrum can create false slander against one another. It can make everyone look untrustworthy to most people. This makes it difficult to know where to look and what news outlets are the better/worse ones.

Shaquiel Pashtunyar
DSC680 Project 2
Milestone 2

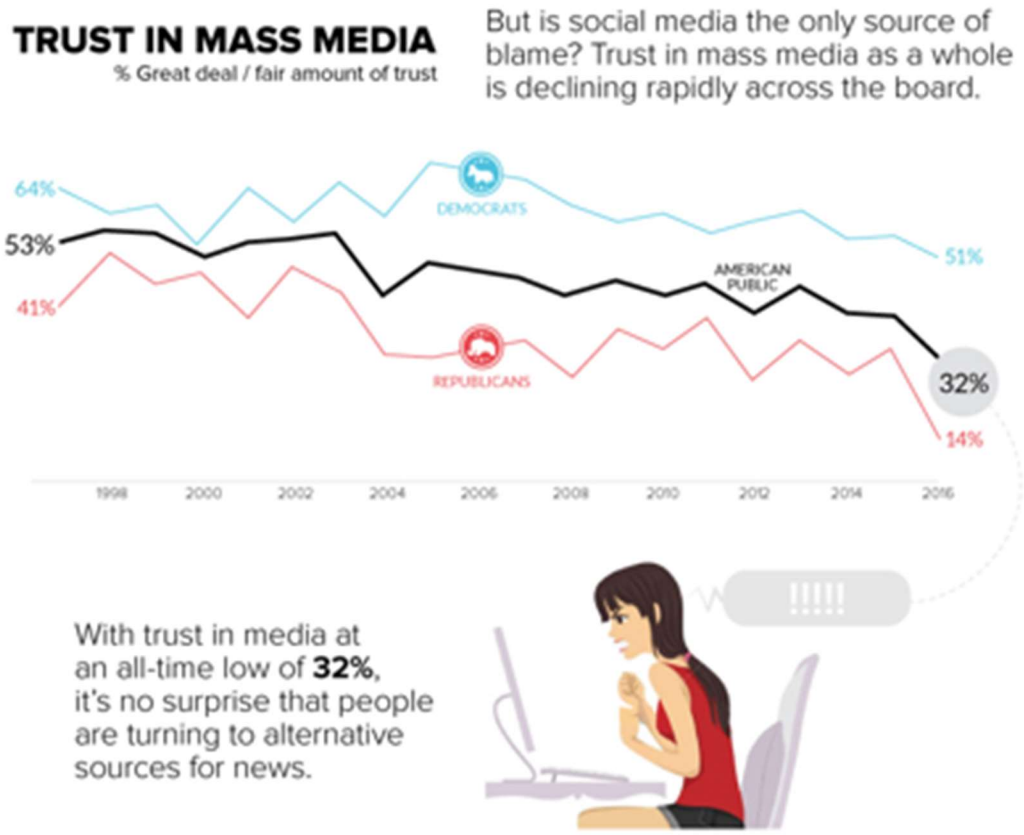

Figure 2 Public trust is at an all-time low with mass media, especially with Republicans in America.

To solve these problems, we will be tackling them all through the lens of solving this media disinformation. We plan to make a tool that will allow users to detect if the article or post they are reading is trustworthy and factual, or if it is fake news. This fake news detection tool will start with an easy implementation of running a few statistical models to help drive you to a trustworthy conclusion of the data.

To begin this experiment, we will need to extract true and false data. The two data sets we will be using are from Kaggle and are a list of True and Fake news articles. The data sets have just four columns of information, title, text, subject, and date. Each data set has around 21-23 thousand articles that we can train an article on, and we will be using these two to create our text predicting tool.

To begin, we will need to do some data munging. The first part of the data munging involves labeling the True and False data sets with a fifth column, and true/false index. This will be Boolean so Fake will be a zero and true will be a one. After we have marked the two, we can combine the data sets into one 'master' table. The 'master' table will contain all five columns, but really, all we need is the index and the text columns. We can drop all the remaining columns to get a target and training column, the text and whether the data is true or false. The final data munging step is to shuffle the data. Since we did simple pandas merge in our code, the first twenty thousand rows are False data, and the second set is true. Because we want our model trained on random datasets, we are going to shuffle the data, so the true and false entries are in no order. This concludes the data munging steps.

Shaquiel Pashtunyar
DSC680 Project 2
Milestone 2

Now that we have a clean and final table, we need to do some additional textual clean up. This is just removing any unique symbols or weblinks that are in the text that could confuse our tools. Now we can move on to splitting the data into our target and training datasets. Luckily, with the fact that we dropped all the other columns in previous steps, we only have two options. Once that is selected, there are several different data modeling tools we can use.

There are a lot of ethical considerations we can start to think about when it comes to predicting if a news article is fake or not. If we do not make the correct prediction, then we are in a way creating more fake news and spreading more disinformation. Due to this, our goal when picking a model is to hover over 95% accuracy. It is also best that we do not just use one model but use several. We want to have high accuracy and precision in our data analysis to make sure that the tool we train is unbiased and others can find trust in it. Because of this, we will train not one model but four of them. The four we will use are logistic regressions, decision tree classification, gradient boosting, and random forest. We want to make sure that they can all have decent results when using the model.

After training all four of the models, we can the score or accuracy of the models. The accuracy is as follows:

- Logistical Regression: 0.9885
- Decision Tree: 0.9965
- Gradient Boosting: 0.9959
- Random Forest: 0.9941

From these results we can see that all four models hit in the 99ish % range which is great news. The decision tree is a little stronger than the others but only by a slim hair of a percentage.

The final operation of our project is to create a tool for users to insert their data. We will do this with a function. The function will leave an open prompt for users to insert information into. They can simply copy the article or text they are reading and feed it into the tool. From there, the tool will analyze the text and give us the response of all four training models. In an ideal world, we would want all four of the models to provide the same answer. This will allow us to have the precision we are looking for to find the tool to be trustworthy.

Shaquiel Pashtunyar
DSC680 Project 2
Milestone 2

```
BRUSSELS (Reuters) - NATO allies on Tuesday welcomed President Donald Trump s decision to commit
more forces to Afghanistan, as part of a new U.S. strategy he said would require more troops and
funding from America s partners. Having run for the White House last year on a pledge to withdra
w swiftly from Afghanistan, Trump reversed course on Monday and promised a stepped-up military c
ampaign against  Taliban insurgents, saying:  Our troops will fight to win .  U.S. officials sai
d he had signed off on plans to send about 4,000 more U.S. troops to add to the roughly 8,400 no
w deployed in Afghanistan. But his speech did not define benchmarks for successfully ending the
war that began with the U.S.-led invasion of Afghanistan in 2001, and which he acknowledged had
required an  extraordinary sacrifice of blood and treasure .  We will ask our NATO allies and g
lobal partners to support our new strategy, with additional troops and funding increases in line
with our own. We are confident they will,  Trump said. That comment signaled he would further in
crease pressure on U.S. partners who have already been jolted by his repeated demands to step up
their contributions to NATO and his description of the alliance as  obsolete  - even though, sin
ce taking office, he has said this is no longer the case. NATO Secretary General Jens Stoltenber
g said in a statement:  NATO remains fully committed to Afghanistan and I am looking forward to
discussing the way ahead with (Defense) Secretary (James) Mattis and our Allies and internationa
l partners.  NATO has 12,000 troops in Afghanistan, and 15 countries have pledged more, Stoltenb
erg said. Britain, a leading NATO member, called the U.S. commitment  very welcome .  In my call
with Secretary Mattis yesterday we agreed that despite the challenges, we have to stay the cours
e in Afghanistan to help build up its fragile democracy and reduce the terrorist threat to the W
est,  Defence Secretary Michael Fallon said. Germany, which has borne the brunt of Trump s criti
cism over  the scale of its defense spending, also welcomed the new U.S. plan.  Our continued co
mmitment is necessary on the path to stabilizing the country,  a government spokeswoman said. In
June, European allies had already pledged more troops but had not given details on numbers, wait
ing for the Trump administration to outline its strategy for the region.Nearly 16 years after th
e U.S.-led invasion - a response to the Sept. 11 attacks which were planned by al Qaeda leader O
sama bin Laden from Afghanistan - the country is still struggling with weak central government a
nd a Taliban insurgency. Trump said he shared the frustration of the American people who were  w
eary of war without victory , but a hasty withdrawal would create a vacuum for groups like Islam
ic State and al Qaeda to fill.


LR Prediction: Not A Fake News
DT Prediction: Not A Fake News
GBC Prediction: Not A Fake News
RFC Prediction: Not A Fake News
```

Figure 3 Sample of the tool inputs, and the result produced from each modeling technique.

Some limitations of the data tool we built are that it only takes textual data. It is not trained on videos and broadcasts that can be a major source of disinformation. This is something we will have to live with because there is a massive challenge in images and spoken news that would need to be accounted for to make the tool work. This tool can still be used with remarkable success, and I can see some web plugins that could run the tool in real time to create analysis and results for individuals quite quickly. The next implementation is to take the tool and make a better UI or plugin so that we can push the tool to have more accessibility to a much wider audience. Regardless, the tool works well, and the results are great. With all four models running with such high confidence, this removes the ethical implications we may have had with the precision and accuracy of the tool!

Shaquiel Pashtunyar
DSC680 Project 2
Milestone 2
10 Questions:

1. What was the training data used in this case?
    a. We trained the model on real and fake news articles release over the last few years.
2. Where was the training data obtained and who obtained it?
    a. It was obtained by Ahmed H, Traore I, and Saad S in a journal article they released to help promote the data set to be used for fake news detections.
3. Who marked each of the news articles in the training data as false or true, how much can we trust them?
    a. The three authors of the data set were the ones who marked each of the datasets for their research article.
4. Are there any other classification tools we can use?
    a. We could also have explored Naïve Bayes, K nearest, vector machine or any other type of classification model. Although that could have been possible, the four we picked seems to give the right amount of variety we need.
5. Where do we see this tool going in the future?
    a. This tool can be used to auto detect websites and news articles with a plug in without having to be run. If we can build that type of functionality into the tool, it can be something with great benefits to society.
6. Is anyone else already using or building these tools?
    a. The data for this tool was release about 4 years ago, there are countless others who have built this tool as well. The question becomes whether their tools are also built into public tools and if people would adopt them.
7. How much will the public be inclined to trust this, or would they rather be in their echo chambers?
    a. This is a tough question to answer because although the tool ahs a high accuracy, ranging around 99%, people tend to trust their own gut to much and the adoption of the tool may be hard to build. If a major tech company had these tools already built into their searches, like FB or Google, I would see the tool gaining traction, but then you get into questionable territory of free speech.
8. How would you solve the problem of videos and broadcasted news?
    a. This would take some development to get there. Right now, we are processing textual data, we would need a tool that takes video speech context and transpose it into something we could plug the tool into. This still leaves the problem of images which we cannot solve at this moment with the current tools.
9. What about visualizations made by an individual that are shared as an image?
    a. This is in the same realm as the previous question. We would need a tool to detect if an image is portraying a reality or a falsehood. This seems a little too advanced for our current tooling.
10. What did you learn from making this tool?
    a. I learned how to make an input command in python, to sort and test a new data set, and how to test several classification models at once. The test of four models at once seems greatly beneficial, and once the data is prepped this is easy to set up and do. It is recommended for a lot of my future work.

Shaquiel Pashtunyar
DSC680 Project 2
Milestone 2
References:
Kaggle dataset that I will be using as the basis of this project:

https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

Yale Article
https://archive-yaleglobal.yale.edu/content/rise-and-rise-fake-news
Fake news graphs: Figure 1 and 2
https://www.visualcapitalist.com/fake-news-problem-one-chart/

Authors of the dataset:

Ahmed H, Traore I, Saad S. “Detecting opinion spams and fake news using text classification”, Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.

Ahmed H, Traore I, Saad S. (2017) “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138).