Mitigating Engagement with False News Stories on Twitter: An Interventionist Approach

Teagan Mucher
Occidental College
Senior Comps Proposal: Department of Computer Science
tmucher@oxy.edu

Fall 2019

Abstract

False news is rampant on social media platforms, affecting our democracy and further deepening fault lines in our society. This proposal describes a system I intend to build for my senior comps project that leverages machine learning techniques and learnings from cognitive psychology and modern disinformation campaigns to mitigate the impact and diffusion of false news on Twitter. I explain the existing body of work around false news classification, and explore the specific ways in which false news is spread on Twitter, especially by malicious actors. I also discuss the methods, tools, and timeline of the project, with the end goal of building a reliable system that can counteract false news on Twitter in real time and at scale.

1 Introduction

In the years following the 2016 US Presidential campaign and in the subsequent search for answers, fingers were pointed at the role that false news on social media played in the election. Social media platforms were designed to spread information quickly and at scale, and as it turns out, falsehoods spread farther, deeper, and quicker than the truth (14). While not all of the blame can be laid at the feet of Facebook and Twitter, unfortunately they were ill-prepared. A sophisticated Russian disinformation campaign weaponized these social media platforms and took advantage of content recommendation algorithms to spread divisive news stories far and wide. Facebook, more so than Twitter, has added various checks and systems for dealing with false news. Despite this progress, the impact false news has on our democracy and political discourse. One area where there is much work to be done is in countering its spread and effectiveness.A personal interest in politics was sparked early on thanks to an ideological father, but it wasn't until the election of our current president that this nascent interest turned into deep fascination. Throughout the last few years, my existing interest in computer science and technology has intertwined with this new attraction to politics. The intersection of the two is incredibly relevant and is a fascinating space currently, especially if you happen to be wonkish about technology and policy.

For my senior comps, I propose a system to mitigate the impact of false news on Twitter through an "antidote" approach. Fighting false news in this way focuses on intervention and the juxtaposition of false news with a sort of fact-check. In this case, the fact-check is not a set of articles that refute the original story, but rather an advisory about the likelihood that a source tweet contains false news. This system will consist of a classification system to detect false news, and a processor for ingressing Twitter data and responding to source tweets. A future component that I have plans to add in the spring will be a system for generating relevant replies. This project is a fairly deep dive into machine learning and natural language processing. False news classification is a deeply challenging and multidimensional problem, and despite all of the work that has been done in this domain, there are few baselines or best practices. Matching replies to the original tweets necessitates a sufficiently complex NLP system that ensures that replies are highly relevant and coherent. Further, all of the 'soft' aspects of this project add challenges. The last thing that I want is a system that backfires and loses the trust of Twitter users before positive impact can be made. Nuance around how we as a society talk about false news, how false news can be classified differently, and the partisan political atmosphere in which we live in move this project beyond the purely technical, and into the domain of politics and public discourse.

2 Background

Fake news is a term that evokes strong feelings from people with ideologies that span the political spectrum. It has taken on a very partisan connotation, with your personal beliefs strongly informing how you define the term fake news. Politicians have co-opted the term to refer to news stories that do not support their policies or are personally are damaging to them. As such, the term fake news has largely lost its original meaning: news that is objectively untrue. In this project, I will follow the lead of many other researchers and only refer to objectively incorrect news stories as false news. False news also encompasses disinformation, misinformation, and rumors. However, I don't want to get tied down to the distinction between news stories that have the intent to deceive, as it is my view that intent does not matter anywhere near as much as the impact of false news.

2.1 The Downsides of our Evolved Advantages

Humans are, unfortunately, deeply flawed organisms. Despite all of our biological advantages that have allowed us to ascend to the top of the food chain, we are poorly equipped to navigate our ever-increasingly complex world. The vast quantity of data and the sheer scale of available knowledge has made us reliant on heuristics – mental shortcuts that allow us to make decisions and solve problems quickly (3) – and intuition. In other words, many decisions we make are based on what feels right to us. We don't have the time nor the resources to thoroughly vet every piece of content we view online.

There are five dimensions along which we evaluate content to determine whether or how much we should trust it (12). The lines between them can be a bit blurred, as they tend to feed off one another. Social media distorts many of these and provides an avenue for manipulation. The good news is that these insights into our cognition also inform a strategy to counter false news.

Social Consensus

We are social creatures, shaped by the actions and beliefs of others. Popularity indicators such as likes or retweets distort our perceptions of truth. Specifically in the context of fake news, these markers signal to viewers that others agreed with this piece of news. This subconsciously leads to people changing their perceptions about the quality of the content (6).

When faced with uncertainty, we turn to social consensus to give us insight into what we should believe. This effect is so profound that often all we ask ourselves is whether or not something sounds familiar or not. Social media distorts how often we hear a claim, thanks to content recommendation algorithms that rank content liked and shared higher than other content. This can lead to people believing information to be true just because they have heard it before, and they've heard it just because others furthered the spread of the information by liking it or sharing it. These are echo chambers, and they become a self-fulfilling prophecy.

Support

We take into account evidence supporting some claim or some piece of information when making a judgement about the truth of something. However, there is a shortcut that humans often take – how much evidence we remember there being for some piece of information. Simple and easily retrievable information, and information that we've heard multiple times, can give us the false feeling that there is more support for a claim than actually exists. This ties to the social consensus phenomena and the following.

Consistency

When assessing a piece of information, we often take into consideration whether it is consistent with our prior expectations and beliefs around what is true or not. We ask ourselves, is this difficult to process, or does it fit into our personal narrative of how the world works?

Coherence

Jonathan Gotschall writes in *The Storytelling Animal: How Stories Make Us Human*, "The storytelling mind is allergic to uncertainty, randomness, and coincidence. It is addicted to meaning" (7). Stories that are narrative, coherent, and logical are far easier to believe and remember. We even go so far to add our own details to a story when remembering it, in order to make it more coherent (12). If a story flows, whether or not that story is actually true, we are more likely to believe it to be true.

Credibility

While we would like to think that we objectively look at the credibility of the source for a piece of information, unfortunately that is not always the case. Once again, we often fall back on intuition – in this case, how do we feel about or towards the source? When this happens, the apparent familiarity of the source is very impactful in our judgement of credibility. Our evaluation of a source of information is also strongly impacted on whether the message is consistent with our beliefs, and if the message is coherent or not.

2.2 The Russian Playbook for Disinformation: False News as a Military Strategy

Regardless of your political affiliation, foreign interference in any democratic process should be deeply troubling. Even if it helps "your side" win, the overall motivation of that foreign actor is to advance their own interests, never the interests of the domestic nation. Foreign interference in something as consequential as an election should be treated like the hostile tactic that it is.

Influencing the perceptions of a population has been a hugely effective and valuable military objective throughout history. Propaganda campaigns have turned neighbors against neighbors, friends against friends, family against family. While old forms of military strength relied on the size of one's army or firepower, the information age has opened new channels of influence and leveled the playing field in that arena. Through an online campaign to spread false news, "Russia is attempting to offset Western technological superiority by going straight to the population and shaping their opinions in favor of Russian objectives…By utilizing the internet as a direct conduit to individual Western citizens, Russia has created an extremely efficient asymmetric weapon" (9).

The tactics Russia employs seek to exploit existing tensions within society. Targeting social fault lines and using tailored messages, debates become drastically more polarized and wedges get driven between us. In doing so, trust rapidly erodes in our society. People stop trusting each other, stop trusting our institutions, and stop trusting the government. This is the end goal of the Kremlin – distort everything so that even the truth stops being objective.

The Internet Research Agency

Through a "troll farm" called the Internet Research Agency, Russian strategists create intricate coordinated campaigns that spread false news through fake accounts and pages on social media. Teams of programmers and content producers manufacture false news with the goal of dividing us and attacking the very idea of truth. It is easiest to understand their tactics through a case study of one of these campaigns.

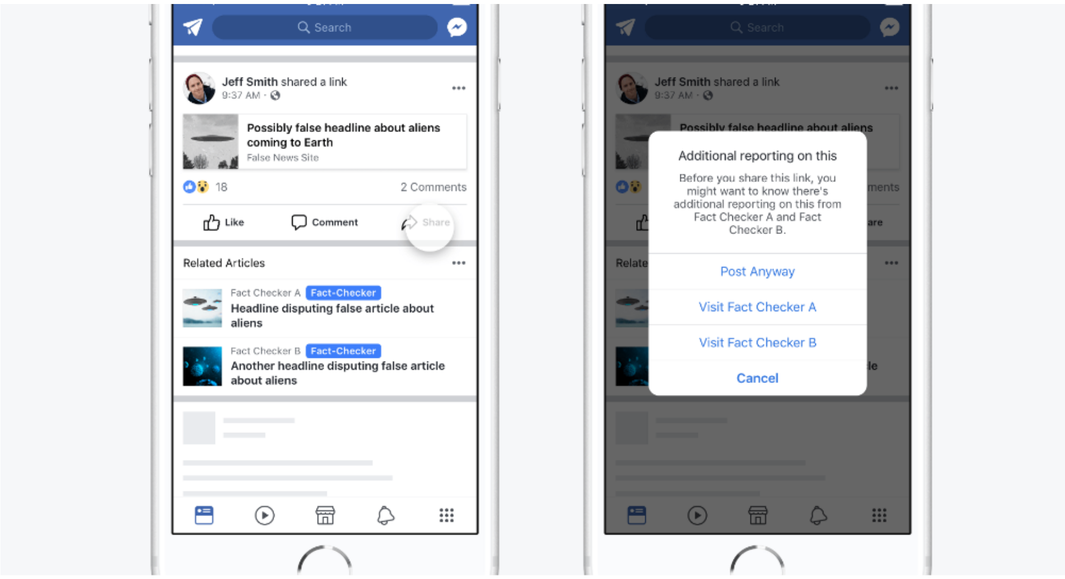
On September 11th, 2014, there was a catastrophic disaster at a chemical plant in St. Mary Parish, Louisiana. Hundreds of Twitter accounts started reporting hearing explosions, and videos spread online showed flames engulfing the factory. Videos on YouTube stated that ISIS had claimed responsibility for the attack, and a screenshot of CNN's website showing a headline reporting on the explosion circulated widely (2).

It was all true and scary and real, until it wasn't. The photos were doctored, the videos purposefully misattributed from other events. But this was no simple hoax, rather a highly coordinated and well-planned disinformation campaign. Certain Twitter users were purposefully targeted to gain maximum attention for the "attack" on the chemical plant. Hundreds of accounts, some of which were bots, promulgated the "evidence". Websites were created that were clones of local news sites, but had headlines describing the chemical plant attack.

2.3 Current Approaches to Fighting False News

Let me be clear, by no means is the Internet Research Agency solely responsible for false news online. Internet Research Agency accounts are rarely the only accounts starting a false news cascade or spreading false news. The false news ecosystem is highly concentrated, and clusters of Twitter accounts regularly link to each other and to the same small set of false news or conspiracy sites (8).

So far, what I have explored has been mostly around what happens on Twitter, and that is



Possibly false headline about aliens coming to Earth
False News Site

Headline disputing false article about aliens

Another headline disputing false article about aliens

Additional reporting on this

Before you share this link, you might want to know there's additional reporting on this from Fact Checker A and Fact Checker B.

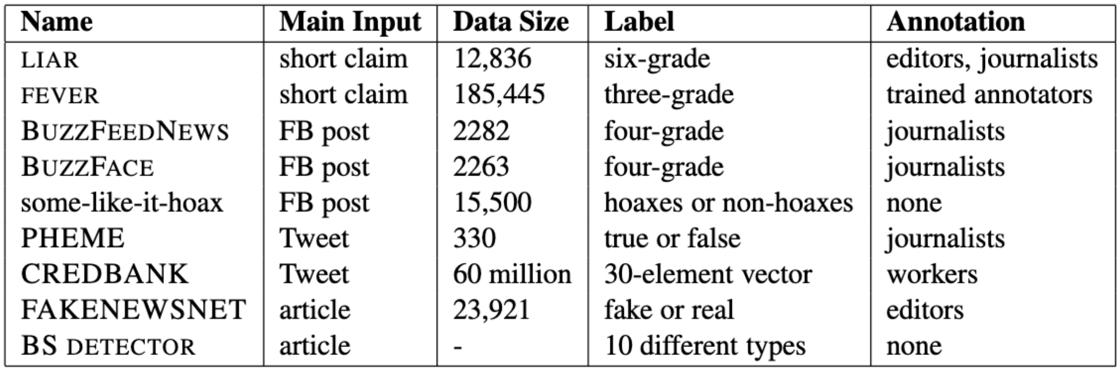
Figure 1: Facebook's current approach to false news

not without reason. Yes, it is very true that inaction on the part of Facebook played a huge role in the ability for false news to proliferate online in the lead up to the 2016 election. However, since then, Facebook has been markedly better in countering false news than Twitter. Employing fact checkers, working with organizations like Politifact and Snopes, and deploying their vast AI capabilities on the problem has had a meaningful impact. "...interactions with false content have fallen sharply on Facebook while continuing to rise on Twitter" (1). While part of their approach has been removing fake accounts and pages, including disabling 2.19 billion accounts in just the first quarter of 2019 (5), an impactful piece of their strategy has consisted of identifying false news and limiting its impact. This is done by a) juxtaposing false news next to articles that fact-check the claim, and b) warning people before they share the disputed story. An example of how this works can be seen in Figure 1.

2.4 False News Classification

False news detection and classification has been a hot area for research in the last few years. Ever since the issues that false news presents became apparent to the general public, research has accelerated – there have been numerous papers published approaching the problem using different datasets, models, and methodologies. Namely, Machine Learning, Data Mining, and Natural Language Processing perspectives have been applied to this problem (10). However, many approaches seem to be limited in their scope, often functioning more like a lab experiment under ideal conditions rather than a practical application. Potthast et al. (11) provide this useful visual framework (Figure 3) for understanding the various paradigms through which false news detection has been researched, including a selection of relevant work. In terms of datasets utilized, researchers utilize a variety of sources, labeling schemes, and annotation strategies to build the sets. Oshikawa, Qian and Wang (10) provide a table summarizing various false news detection related datasets (Figure 2).

The LIAR dataset (15) was the first large benchmark study compiled for false news detection. This data set is comprised of 12.8K short statements and their truthfulness rating from PolitiFact.com, as well as the speaker of the statement and the context in which it was said or posted. Further, there is a justification for the truthfulness decisions and the sources used. The downsides of this dataset are that the unstruc-



Name	Main Input	Data Size	Label	Annotation
LIAR	short claim	12,836	six-grade	editors, journalists
FEVER	short claim	185,445	three-grade	trained annotators
BuzzFeedNews	FB post	2282	four-grade	journalists
BuzzFace	FB post	2263	four-grade	journalists
some-like-it-hoax	FB post	15,500	hoaxes or non-hoaxes	none
PHEME	Tweet	330	true or false	journalists
CREDBANK	Tweet	60 million	30-element vector	workers
FAKENEWSNET	article	23,921	fake or real	editors
BS detector	article	-	10 different types	none

Figure 2: Existing false news detection-related databases

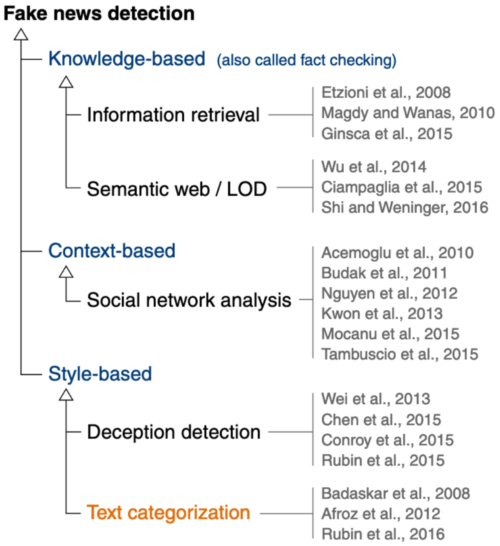


Figure 3: A framework for understanding approaches to false news detection, with selections of relevant work

tured and variable nature of the statements has made training accurate classification models on it non-trivial. Wang (15) saw the best results with a hybrid CNN integrating both text and metadata, but yet only managed to get at best 27.4% accuracy on a test set.

The FEVER dataset, provided by Thorne et al (13), contains over 180k claims taken from Wikipedia data, each labeled as Supported, Refuted, or Not Enough Info. This dataset and original study are more focused on the ground truth of a short claim and setting it up for future use in information verification tasks.

The BuzzFeedNews dataset collected posts from 9 news agencies on Facebook, from both left and right-leaning sources. BuzzFace extends this dataset by including 1.6 million comments on the original articles. Also gathered from Facebook, some-like-it-hoax takes a similarly balanced approach and consists of over 15k posts from 14 conspiracy and 18 scientific organization Facebook pages.

PHEME and CREDBANK would seem like two of the more relevant datasets for what I propose to do. PHEME is focused on rumor-spreading on Twitter, and consists of threads of 9 newsworthy events with true-or-false labels. CREDBANK covers far more events – 1,049 – labeled along a 30-dimensional vector for truthfulness by Amazon Mechanical Turk annotators.

FAKENEWSNET is a more recent dataset containing news content, social context, and spatiotemporal information. Ground truth labels of fake or real are gathered from 2 fact-checking websites: PolitiFact and GossipCop (a celebrity gossip debunking site). The other entire-article dataset, BS Detector, is data collected from the browser extension BS Detector. The way it classifies false news is by searching "all links on a web page at issue for references to unreliable sources by checking against a manually compiled list of unreliable domains" (10).

For most of the research around these datasets, models built have a range of precision and recall, but more importantly, aren't tested in real environments. Further many of the datasets use past news events. Therefore, to use these datasets for classifying false news in the wild, I believe that one needs to use features that aren't specific to the news events themselves, but rather everything around that: what sites tend to publish false news, sentiment and syntactic analysis of the tweet bodies, user/cluster behavior patterns, and so on.

3 Methods

The majority of research around false news classification has been directed towards false news in general, not specifically false news on Twitter. One characteristic that many false news-propogating accounts share is that they include links to ostensibly back up the claim that they are making. Many of these sites they link to are conspiratorial, to put it mildly. I propose to start with a false news classifier and the data pipeline for pulling tweets containing links to unreliable sites. I then will layer on additional models to improve the quality of the probability distribution that a tweet is spreading false news. These additional models will vary in what they analyze – one idea is around semantic and syntactic analysis of the news story that is posted. This project is ambitious in its scope, both in the quantity of components I want to build, and the technical complexity of each of those components. The false news classifier and the natural language processing component draw on modern machine learning techniques. Fortunately, much work has already been done around classifying false news, and I intend to build a classifier very similar to existing work. My experience with Python and familiarity with machine learning algorithms will be very useful throughout this project. For development, I will be using Python and a few libraries (Tweepy, Pandas, scikit-learn, and TensorFlow) and Jupyter Notebooks. Using these tools inside of a virtual environment will make my project easily reproducible. Jupyter Notebooks will be used for visualizing data and prototyping the classifier, while some Python scripts will be standalone. The architecture of the system can be found below.

3.1 Timeline

Milestone 1: End of October '19

The goal in the first milestone is to build a simplistic model for predicting likelihood of false news given a URL. A full-fledged false news classifier isn't in scope for this milestone. I will start with the FAKENEWSNET dataset, using the PolitiFact data. First, I need to manipulate the datasets and pull the urls from their .csv files, and add a label given which file they came from. Preprocessing of the data consists of removing extraneous information from the URLs: namely, anything that isn't just the domain. I will experiment with a few machine learning methods for making predictions. From each model, I will evaluate the F1 score, picking the best model after implementing them all.

The first method I will try is logistic regression, using the LogisticRegression class from Scikit-Learn. Logistic regression is a machine learning model that works well with modeling a binary dependent variable, which in my case is false/real. In other words, given an instance of a class, predict the probability that the given instance belongs to a particular class. After logistic regression, I will then try support vector machines, specifically the regression-oriented variants from Scikit-Learn: Linear Support Vector Regression and Epsilon-Support Vector Regression. Support vector machines are quite versatile models that can work well on small- or medium-sized datasets, like FAKENEWSNET. I will try the two variants I selected due to their known efficacy in both linear and non-linear regression tasks.

Third, I will try variants of neural networks.The simplest type of artificial neural network is the Perceptron. However, Perceptrons have their limitations, namely that they do not output a class probability. I can overcome this by using a Multi-

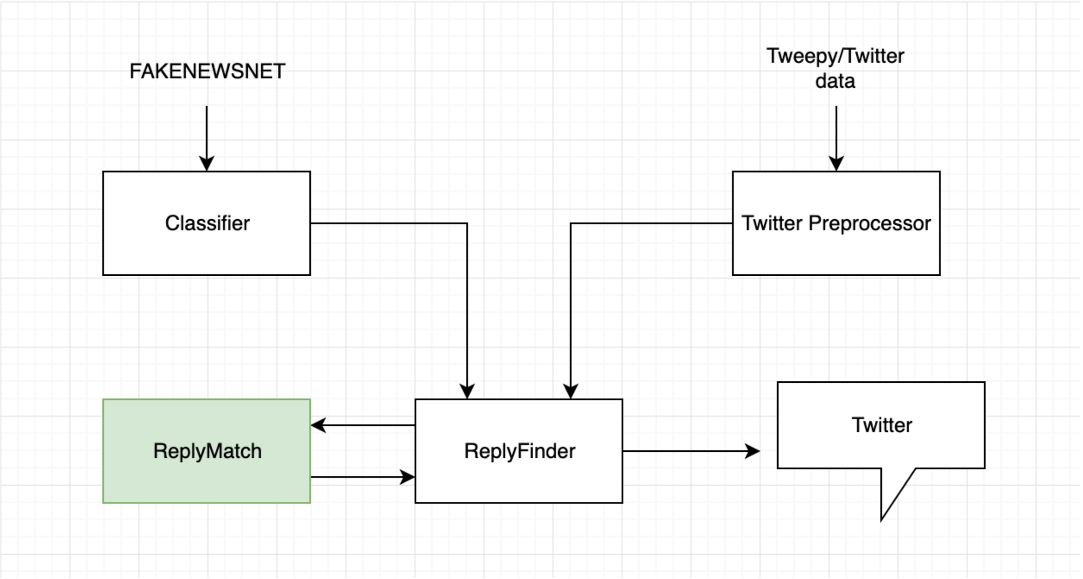


Figure 4: High-level architecture

Layer Perceptron with a final softmax layer to predict probabilities. This is known as a Feed-forward Neural Network. The inputs will be the URLs, first trying just raw text and treating each URL like categorical data. For implementing a neural network, I will try scikit-learn's MLP-Classifier method, and then turn to TensorFlow. Using the TF.Learn API, I can train a neural network with various hidden layers and a specific number of softmax output neurons (in my case, 2). I will do what is possible to tune hyperparameters, trying GridSearchCV for scikit-learn and potentially the Tune library with Tensorflow.

Milestone 2: Mid-November '19

In this milestone, I will build the full end-to-end system to crawl for tweets and respond. Looking for tweets containing URLs found in the FAKENEWSNET dataset, I will use those URLs as inputs into my classification model. I will use the Tweepy library for Python to access the Twitter API. Setting up a StreamListener, I will listen to all tweets being posted on the platform in real time, filtering based on the full set of URLs from both the real and fake CSV files. For each tweet, the URL will use the same preprocessing method that is used for the classifier data. The URL from the received tweet will then be tested using the classifier, returning a probability that the news story from the given URL is false.

The second component to this system is the response functionality. I will use the Tweepy update_status method to respond to the original tweet. This is a tricky part to get right, as the language I use as templates for responses should be unbiased and focused purely on the objectivity of the false news assessment. I will seek guidance from my academic and comps advisor on the specific language used here. It could be fairly easy for my response tweet to backfire, and further entrench someone's belief in a false story. Thus, the response needs to come off as friendly and not challenging or attacking someone's beliefs, but rather solely assessing the veracity of a news story. The reply tweet will most likely include the probability of false news, given the URL, essentially just focused on objectively informing the original tweeter.

I propose to measure the efficacy of my system by randomly grouping a source tweet into a control group or experimental group, and then seeing if aggregate engagement has a statistically significant difference between the two groups. The experimental group will receive a response with an intervention tweet, while the control group will not. Tracking engagement per tweet won't give me meaningful information, as the impact of a story can vary widely. Thus, I will compare aggregate engagement between the two groups.

Milestone 3: End of November '19

This milestone is all about enhancements to the classification model and the efficacy of the system. I can continue to use the FAKENEWSNET. The first idea I will try is around weighting certain users, given their propensity to deceive. A sort of credibility-rating like this could be very useful. The first approach I would try here is, given a tweet with a URL from my set, analyze all of their tweets containing links from my set. This proportion of false:true can be used as a credibility score. I can analyze subsets of users, finding their credibility scores and building sample distributions. The average mean of the distributions would inform the point at which a credibility score would weigh up or down the original prediction from the classification model. The idea here is that it's likely that non-malicious actors on Twitter are likely to spread something false at least a small percentage of the time, even if they don't mean to. The mean of sample distributions would provide a threshold.

The second approach, while not machine learning focused, would consist of identifying subsets of users to focus on with the goal of increasing impact on total engagement. A large-scale study from the Knight Foundation (8), among others, could inform the specifics of this approach, but related work around this topic suggests that the spread of false news on Twitter is very centralized and clustered. For example, just focusing on the 10 largest accounts that regularly spread false news and targeting the original tweet and the total tweet cascade, I believe that I can make an impact on mitigating engagement with false stories. I propose to measure the impact of this change first by randomly grouping a source tweet into a control group or experimental group, and then seeing if aggregate engagement has a statistically significant difference between the two groups. Meaning, whether or not I respond down the tweet cascade or not, does my intervention affect total engagement.

If I have time in this milestone, I will try to add another layer to the summary model. Another related idea around network-based approaches makes use of a structured knowledge network called DBpedia. Conroy, Rubin and Chen (4) propose a method using a shortest-path finding algorithm to determine the veracity of a claim. The closer the nodes, the higher the semantic similarity, the higher probability that a "subject predicate-object statement is true" (4). I would attempt to implement this method, layering the resulting probability into my summary model.

4 Discussion

Political discourse around false news presents a challenge for this project. Attempting to counter the impact of false news can easily be thought of as an attack on one side or the other, depending on the reader's personal views. I strive towards objectivity and truth, and will make a consistent and ongoing effort to stop any of my own political views from creeping into my methodology or analysis. However, that is not to say that each side will be equally represented. Russia targeted for the most part conservatives or Republicans, and current datasets for training models on are imperfect.

On the technical side, it is hard to ever have complete certainty that my model is classifying correctly. Confidence in my model predictions should increase with the various enhancements I discussed, but I believe that transparency of the probability (in the response tweet) is important for Twitter users who encounter my system. Human raters don't always agree on the veracity of a story, and reducing the dimension of labels to just two – false and true – abstracts away nuance that can be important. However, I believe for the purposes of this project, simplification is a necessary evil.

References

[1] Hunt Allcott, Mathew Gentzkow, and Chuan Yu. Trends in the diffusion of misinformation on social media, 2018.

[2] Adrian Chen. The agency, 2015.

[3] Kendra Cherry. How heuristics help you make quick decisions or biases, Jun 2019.

[4] Nadia Conroy, Victoria Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. In *Automatic Deception Detection: Methods for Finding Fake News*, 10 2015.

[5] Lauren Feiner and Salvador Rodriguez. Mark zuckerberg: Facebook spends more on safety than twitter's whole revenue for the year, 2019.

[6] Center for Information Technology and Society at UC Santa Barbara. How is fake news spread? bots, people like you, trolls, and microtargeting.

[7] Jonathan Gotschal. *The Storytelling Animal: How Stories Make Us Human.* Mariner Books, 2013.

[8] Matthew Hindman and Vlad Barash. Disinformation, 'fake news', and influence campaigns on twitter, 2018.

[9] Timothy P. McGeehan. Countering russian disinformation. *Parameters*, 48(1):49 – 57, 2018.

[10] Ray Oshikawa, Jing Qian, and William Yang Wang. A survey on natural language processing for fake news detection. *ArXiv*, abs/1811.00770, 2018.

[11] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 231–240. Association for Computational Linguistics, July 2018.

[12] Norbert Schwarz, Eryn Newman, and William Leach. Making the truth stick & the myths fade: Lessons from cognitive psychology, Feb 2017.

[13] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *ArXiv*, abs/1803.05355, 2018.

[14] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 09 Mar 2018.

[15] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *Association for Computational Linguistics*, 2017.