Identifying Fake News in Social Media Content

Katon Minhas

Introduction:

Misinformation has historically been acknowledged as one of the most effective methods for creating chaos, confusion, and division in a population. With consequences ranging from defamation of an individual's character to bolstering the platforms of autocrats, fake news has proven to be a constant source of concern among journalists, political scientists, and the general public. It is only recently however that the issue became a relevant topic of research for linguists, social media experts and leaders in the artificial intelligence community. While misinformation has always maintained a presence in obscure internet forums, the emergence of social media platforms such as Facebook and Twitter as leading news sources have drawn the problem to the mainstream. The volume and severity of misinformation on social media was made apparent in the 2016 E.U. membership referendum (commonly known as Brexit) and the 2016 United States presidential election. From February 2016 to November 3rd, engagement on the top 20 election-related stories from mainstream news outlets decreased from 12 million to 7.3 million engagements on Facebook. Over that same period, engagement on the top 20 fake news stories nearly tripled, rising from 3 million to 8.7 million (Silverman 2016). There is some evidence to suggest that the problem has only gotten worse since then (Rogers 2020). This unprecedented rise in misinformation has since triggered a worldwide reckoning on the potential of social media to erode democracy and contribute to societal division.

As a part of this continued discourse, the term "fake news" has emerged as the popular catch-all phrase to describe any "false, inaccurate, or misleading online information, designed, presented, and promoted with malicious intent or for profit" (ENISA 2018). Some scholars maintain that hyper-partisan political opinion programming should be considered fake news as well; pundits who employ calculated omission of information in conjunction with heavily biased commentary technically do not present falsehoods, but their overall conclusions and implications largely align with those of traditional fake news (Herrman 2016, Benkler 2017, Holt 2019). A wide range of stories and statements fall under this umbrella. PolitiFact, an esteemed fact-checking project operated by the Poynter Institute for Media Studies, rates statements on a scale of "True" (statement is accurate, with no notable omissions) to "Pants on Fire" (statement makes a ridiculous claim) (PolitiFact). As of May 2021, less than 30% of statements checked were rated "True" or "Mostly True", while "False" had the most ratings with 21%. Unfortunately, the statements checked on PolitiFact are only the tip of the iceberg; the sheer scale of fabricated news stories and false statements by politicians has become an overwhelming barrier for fact checking journalists.

There are a number of factors contributing to the recent rise in online fake news. One reason is that fake news is generally much faster, easier, and cheaper to produce than legitimate news (Shu et al, 2017). Traditional news production is time consuming, requiring journalists to verify sources and adhere to a consistent publication standard. Fake news inherently doesn't operate under these constraints, allowing for a much higher daily output from a single production source. Additionally, traditional news outlets are very costly to run – elite journalists demand high salaries and websites or TV stations are expensive to operate. Fake news content can be cheaply disseminated via a social media account or blog.

Overreliance on social media is another significant contributor to the rise in fake news. In 2018, roughly two-thirds of Americans got at least some of their news from social media – more than both print and television (Matsa and Shearer 2018). Numerous studies have found that social media often limits the diversity of sources and opinions users are exposed to (Cinelli et al, 2021). Lack of exposure to contradictory information allows fake news statements to be seen as unchallenged, resulting in more false or biased beliefs to be perpetuated as normal. Additionally, the networked nature of social media allows for frequent virality of stories, with fake news stories more likely to spread faster and farther through the network than real news stories (Vosoughi et al 2018). One reason for this spread is the novelty of a fake news story when compared to a real news story. News regarding novel events ("Breaking News") spreads faster due to the perceived urgency surrounding it. Since fake news stories frequently report on events that never happened, they are generally interpreted as more novel than real news stories, and spread faster as a result. Lastly, fake news tends to produce strong negative emotions. In a controlled study, fake news stories were rated higher in measures of fear and disgust, while real news stories were rated higher in joy, trust, and anticipation (Vousoughi et al 2018). In short, fake news relies on outrage and divisiveness to facilitate a fast spread throughout social networks.

Further exacerbating this issue, studies have shown that social media users display a "stunning and dismaying consistency" in their inability to detect fake news online (Wineburg et al 2016). Despite their stereotypical internet literacy, only 20% of middle school, high school, and college students were able to successfully discriminate between a verified mainstream news source and a fake news source, with 30% of the students actually arguing that the fake news source was more trustworthy. Additionally, 80% of students believed that a paid ad was a real news story, and the same percentage were willing to accept an image with a misleading caption as fact without asking any questions or attempting to verify the information.

Skyrocketing production, quick spread, and the ability to evade human detection are all factors that have elevated the fake news phenomenon to the scale of a serious crisis. Fake news is now considered an established threat to democratic principles, civil discourse, and the existence of social media ecosystems as they exist today. The severity of the issue has inspired a surge in research surrounding fake news and fake news detection.

Thus far, the majority of automated fake news detection systems utilize one or more of the following four perspectives (Zhou and Zafarini 2020):

1) Knowledge-based – automated fact-checking systems that assess a claim's authenticity by comparing its content and implications with previously verified facts.
2) Style-based – natural language processing (NLP) systems that extract and assess distinguishing stylistic features of the text of the claim. It is found that fake news is more informal, subjective, emotional, and profane than real news (Zhou and Zafarini 2019).
3) Propagation-based –graph theory-inspired approaches which examine patterns in the way news through the social network. It is found that fake news is spread farther, spread by more users, and generates stronger engagement that real news (Mustafaraj and Metaxas 2017). Additionally, the networks of fake news spreaders are significantly higher density than those of real news spreaders (Zhou and Zafarini 2019). While propagation-based methods have proven effective, they are not considered an effective real-time solution because they require the news to have already spread through the network.

4) Source-based – systems that assess the credibility of the source of fake news based on assessment of that source's history. Sources include writers, publishers, and spreaders of the news.

Decomposition

The goal of this project is to present a feasible, reliable system for detecting fake news text on social media sites. As the topic of fake news is highly politicized and controversial, an effective system must be overly transparent and work to minimize bias wherever possible. The motivations behind this project are two-fold: 1) The rise of fake news has measurable negative societal consequences, and 2) As legislative pressure to regulate social media content increases, user-generated fake news poses a major threat to the existence of social media in its current state.

While the harmful effects of fake news were previously limited to gradual and quiet shifts in attitudes, the COVID-19 Pandemic, the 2020 U.S. presidential election, and the increase in foreign manipulation on social media have led to more direct and measurable links. Misinformation surrounding the COVID-19 pandemic is particularly extreme and widespread. One survey in March of 2020 found that nearly 30% of U.S. adults believed that the coronavirus was a Chinese-made bioweapon, while another found that 85% of U.S. adults believed that at least one COVID-19 conspiracy theory was "probably" or "definitely" true (Romer 2020, Miller et al 2020). Misinformation surrounding the pandemic has been linked to a myriad of negative effects, including increases in violent attacks against doctors and Asians (Said 2020, Levin 2021). These patterns are not just seen in the United States – pandemic-related misinformation has taken hold in communities around the world (Taylor 2020, Gharib 2020). Similarly, the "epidemic of malicious fake news" surrounding the legitimacy of the 2020 presidential election has been described as a threat to democracy, both in the U.S. and abroad (Zengerle 2016). A January 2021 poll found that one-third of Americans (and two-thirds of Republicans) adhere to the idea that Joe Biden's victory is illegitimate (Bump 2021). Falsely-based concern over election security has led to numerous poll workers and election officials reporting harassment, assault, and death threats (Gstalter 2020, Lee and Brown 2020). Due to its remarkable efficacy as a creator of chaos, violence, and division, fake news has turned into the weapon of choice for foreign interests. In a 2017 hearing before the Senate judiciary committee, Facebook CEO Mark Zuckerberg testified that Russian-backed fake news posts directly reached 29 million Americans, and indirectly reached as many as 126 million (Zuckerberg 2018). If left unchecked, false and foreign social media propaganda will likely remain an effective weapon. The overall destructive effects of fake news on social media cannot be understated.

In addition to harmful societal effects, fake news may threaten the very existence of the social media sites they make use of. Section 230 of the Communications Decency Act currently provides immunity from liability for internet-based providers who host information originating from third-party users (47 U.S.C. sec. 230). However, since 2016 the law has come under increased scrutiny from both liberal and conservative policy makers. Democrats maintain that the law allows social media sites to turn into breeding grounds for hate speech, extremism, and misinformation. Republicans argue that content-moderation on social media is biased against conservative views while protecting their liberal counterparts. A repeal of Section 230 would make social media sites responsible for the content they host and open them up to potentially massive lawsuits. Dominion Voting Systems is currently in the midst of several of multi-billion-dollar defamation lawsuits against Fox News and Newsmax, among many other parties, for their roles in spreading the falsehood that the company's voting machines were designed to commit election fraud. A repeal of Section 230 would open Twitter, Facebook, and other social media sites to an even more overwhelming level of litigation (Allyn 2020). In that event, the inability of social media companies to comprehensively review potentially litigious content could very likely eliminate social media as we know it.

For this project, I am proposing a style-based fake news detection system. The system will take features extracted from the text of a social media post and output a prediction of the post's veracity. The model will be a deep learning model with parameters and hyperparameters fine-tuned through experimentation. It will be trained using data labelled as fake or real by a reliable fact-checking organization. Once the model is developed and sufficiently trained using the labelled dataset, it will be tested further to evaluate performance in the desired social media environment (likely Twitter or Facebook). A preliminary method of classification will be required at this point to determine if a given social media post is meant to be taken as news – only posts that attempt to pass themselves off as factual, or contain opinions that obscure significant facts, will be fact-checked. A team of journalists will act as fact-checkers to supervise the model and correct mistakes, while any persisting errors will be corrected by the development team. Concurrently with this process, cognitive psychology and user-interface experts will research and test hypothetical final products to determine the correct user-facing presentation of the model results on the social media site. When the model is ready to be fully deployed, continuous monitoring of the model performance and its efficacy in preventing the spread of fake news (via network analysis). Continued adjustments will be made. At this stage, the model is working as part of a human-centered AI system; final fact-checking decisions are still made by journalists, with the model acting as a recommender to speed up the fact-checking process. As continued improvements are made, the journalistic team will shrink and assume more of a supervisory role, correcting errors where needed but for the most part allowing the system to detect fake news in a nearly autonomous fashion. It is likely that the model will unable to improve past this stage to reach a level where it can work fully autonomously.

Domain Expertise

Traditionally, the hard task of taking on fake news has fallen to journalists. Major fact-checking journalistic organizations include PolitiFact, FactCheck.org, FlackCheck.org, OpenSecrets.org, the Washington Post, and Snopes. These organizations employ teams of journalists and editors to review a selection of articles and statements and make a determination of the degree of their veracity. Each group also operates their own social media accounts to report significant findings directly to user's Facebook and Twitter feeds. Unfortunately, journalistic fact-checking operations remain largely unautomated, which has limited their effectiveness as the volume of fake news continues to rise. Additionally, while great care is taken to minimize bias, the system is not perfect and some bias is inevitable. An example of this the process of choosing statements to fact-check. PolitiFact and FactCheck.org select statements by manual review, or by taking suggestions from readers. This means that an enormous number of statements are left unchecked simply because they are not selected for checking.

Despite their limitations, fact-checking journalists will likely remain an integral part of an automated fake news detection system. Any style-based or propagation-based model will not be able to achieve perfect 100% accuracy, and false-positives or false-negatives can be highly consequential in certain circumstances. A high-profile claim being incorrectly labelled true or false could potentially have devastating political or economic effects. For example, after 2013 hack of the Associated Press Twitter account resulted in a tweet claiming that President Obama was seriously injured in an explosion, an estimated $130 billion in stock value was lost (Rapoza 2017). Already devastating fake news can be made worse by incorrect and unchecked model performance. Moreover, even in the case that a model is developed with perfect performance, human fact-checkers will still be required for the purposes of transparency and to build trust in the system. Especially in systems relying on non-fact-checking techniques (style-based, propagation-based, or source-based), having a team of qualified journalists verify classifications will assure near-perfect accuracy and build faith in a populace that can be untrusting of algorithms.

In addition to journalists, social media experts and network analysts could prove useful for their advanced and detailed understanding of information propagation through networks. Since the explosion in popularity of Facebook and Twitter in 2008, research on social networks has greatly increased. Much of the research has been "devoted to the analysis of the propagation of information, influence, innovations, infections, practices, and customs through networks" (Chen et al 2013). This field of knowledge is very relevant to the problem of fake news detection and intervention, especially in the case of propagation-based classifiers. By examining misinformation from an epidemiological viewpoint, social network researchers have essentially developed a more advanced and mathematics-based version of the news cascades commonly seen in more traditional journalism. While journalists tend to view the spread of news as the transmission of a single unit (the story or article) to a set of isolated readers, it is now more appropriate to think of news as a dynamic cascading stream that is subject to constant evaluation, criticism, and change – narratives are increasingly subject to interpretation and outside influence by non-experts (Langeveld 2009).

The last and perhaps most crucial group of domain experts required by this project are cognitive and social psychologists. Different instances of online fake news present a diverse array of cognitive hurdles standing in the way of a successful intervention. Chief among these is the issue of confirmation bias – people tend to accept information that matches their previously held beliefs, while tuning out and minimizing information that challenges them (Koriat et al 1980). Confirmation bias is considered to be the driving force behind the echo chamber effect – when people exclusively see confirming information, the bias grows stronger and they are less increasingly less likely to have their mind changed. EEG studies have shown that social media users exhibit significantly greater cognitive activity when reading news headlines that align with their own political believes, while effectively ignoring headlines that challenge their opinions (Moravec et al 2018). Moreover, placing a fake news flag on headlines that align with users' opinions had no influence on users' belief that the headlines were true. This finding presents a significant challenge – perfect model performance is rendered useless if users choose to ignore the label. Some research suggests that appropriate interventions are in fact able to "alleviate the negative effects of selective information processing on issues such as political polarization" (Rollwage and Fleming 2021). This is accomplished by 'wearing down' the confirmation bias effect – observing that one's beliefs are considered false on multiple occasions over time may weaken confirmation bias and make the individual more willing to accept contradictory information. Such an effect would not be apparent in the previously mentioned EEG study because the study was limited to one controlled session. Furthermore, a 2019 study revealed that the type of warning label used significantly affected social media users' willingness to accept news as fake. It was found that participants' detection of fake and real news improved when fact-checking warnings were presented, while a machine-learning-graph warning increased participants' sensitivity in differentiating fake news from real news (although it did not increase their trust in the ML model) (Seo et al 2019). The results of these studies indicate that the consistent presence of a carefully optimized warning label will likely be effective in decreasing belief in fake news, while also increasing users' ability to recognize fake news on their own. Further research in this area is required to determine exactly what the most effective style of warning label would look like in practice. Due to the delicate psychological factors that influence belief in fake news, cognitive psychologists will play a vital role in the development process, particularly in the design of the front-end user interface.

Data

Several Class A datasets have been compiled for the purposes of machine-learning for fake news detection (Lawrence 2017). The University of Victoria information security and object technology (ISOT) research lab created a dataset with more than 60,000 full articles classified as "Fake" or "True" by PolitiFact (Ahmed et al 2017, Ahmed et al 2018). The articles are dated from 2016 to 2017 and contain primarily world news and political news. Another commonly cited dataset is the LIAR dataset, compiled by researchers at UC Santa Barbara (Wang 2017). The paper also uses PolitiFact but instead focuses on short statements (in the form of speech clips, interviews, written releases, etc.) from political figures, as opposed to full length news articles. A third popular data repository is FakeNewsNet, which uses data scraped from PolitiFact (for political stories) and GossipCop (for celebrity gossip and entertainment) (Shu et al 2020). Several other lesser-used datasets exist, most of which also make use of PolitiFact fact-checking and API to curate their data. The majority of existing models were developed using one or more of these datasets for training.

A key data collection and application development tool is Twitter's Tweepy library, an open-source Python library that allows developers to access the Twitter API. The library allows users to create applications and bots for Twitter, providing search and access to a continuous real-time stream of Tweets. The API also provides access to followers/following data. This provides a graph-like representation of a local area of the greater social network, which can be used to map the spread of fake news for propagation-based systems.

Design

Thus far there have been many attempts to develop machine-learning or deep learning classification systems for fake news detection, achieving varying degrees of success. Style-based classifiers relying on NLP techniques to extract textual information remain very popular. de Oliveira et al (2020) achieved 86% accuracy on a database of Tweets using a one-class support vector machine. Ksieniewicz et al (2020) used a variety of methods including bayes classifier, multi-layer perceptron, and Hoeffding tree to achieve peak classification accuracy of 81%. Tacchini et al (2017) opted for a source-based approach to detect hoaxes (particularly absurd and conspiratorial fake news posts) by taking the quantity and level of community overlap of likes on Facebook posts. Performing logistic regression on the "Like" data resulted in 99% classification accuracy, indicating that the set of users who interact with social media posts is highly indicative of the veracity of that post's content. Lastly, Reis et al (2019) used a hybrid supervised learning approach, taking stylistic textual features, source credibility, and propagation data. Several classifiers were used, with random forest (85%) and gradient boosted decision tree (86%) achieving the best performance.

While source-based, propagation-based, and hybrid classifiers are commendable for their high performance, they are likely limited in their practical use. Source-based classifiers work under the assumption that news published by a source that has produced fake news in the past is more likely to be fake than is content produced by reputable sources. While this assertion may be true, it comes with the caveat that any misclassifications (e.g., a piece of legitimate news published by an unreputable source being classified as fake) are based solely on the publisher and not on the merits of the news itself. This means that models based on publisher-credibility will rightfully be accused of being biased. The successful source-based approach by Tacchini et al (2017) forgoes emphasis on the publisher in favor of examining the groups of users who interact with the post. While this approach was highly successful and remains a viable method for retroactive fake news detection, it is not practical as a method for real-time detection. Since the model relies on user interaction to make its prediction, it can only make a determination after a large set of users have seen, interacted with, and spread the post. By this time, the damage will have been done. A similar effect exists for propagation-based methods – models that rely on propagation trees are only effective after the news has sufficiently spread through the network. For these reasons, style-based classifiers have been the popular choice among researchers in this area, despite their performing with equal or less success than other methods.

With these restrictions in mind, I developed a simple neural network model using stylistic textual features extracted from the University of Victoria ISOT dataset. Both headline features, article text features, and a combination of the two were used. The four features included were:

1) Polarity – A measure of how polarizing the position taken by the text is, using positive or negative sentiment as an indicator. Fake news is found to be consistently more polarizing than real news, as well as more skewed towards negative positions. The TextBlob sentiment analysis polarity function was used to produce the scores. Scores range from -1 to 1, where a value of 0 indicates a completely neutral position and a value of -1 or 1 indicates a completely negative or positive view, respectively.
2) Subjectivity – A measure of the level of opinions versus facts expressed in the text. Fake news is consistently more opinion-based than real news. Using TextBlob subjectivity function, scores were produced in the range of 0 to 1, where a score of 1 indicates an entirely opinion-based statement and a score of 0 is purely factual.
3) Offensiveness (Informality & Emotionality) – Indirectly measured by the strength and quantity of profanity in the text. Real news text hardly ever contains profanity, except in the context of a quote, while fake news text may or may not use it frequently. The Python profanity-check library was used to generate scores between 0 and 1, where the score indicates the probability that the text contains offensive language. Level of profanity is highly correlated with emotionality and informality; for this reason, I used profanity as a proxy for both, combined into the property "Offensiveness".
4) Word Diversity – A measure of the percentage of unique words in a text. Fake news typically uses more diverse language while real news tends to remain prosaic and concise.

Using the four above features as input data, I developed a simple Keras sequential model. The model consisted of a dense input layer and 2 dense hidden layers, all using ReLU activation.

The Adam optimizer was used, and I experimented with different learning rates between 0.001 and 0.005. The model was trained using 50 epochs on an 80-20 train-test split of the data.

In addition to the Keras model, I created an AutoML run for the headline and the article data, both with and without principal component analysis (PCA) preprocessing. The best-performing model for the raw input headline data was MaxAbsScaler, LightGBM (light gradient boosting machine), while the best model for the PCA input headline data was StandardScalerWrapper, XGBoostClassifier. The MaxAbsScaler, LightGBM model also performed best for the combined article and headline features.

Diagnosis

The Keras model had modest success when only using features generated from the article headlines, classifying 72.31% of headlines successfully. A slightly higher average success rate of 75.45% was achieved using features from the entire article text. The model performed best when the four features from both headlines and article text were included (8 features total), achieving 81.3% average classification accuracy. Experiments with PCA preprocessing on the raw feature data as a feature extraction technique and for dimensionality reduction did not affect the classification accuracy in any case.

The MaxAbsScaler, LightGBM model classified the raw headline feature data with 73.6% accuracy (AUC macro = 0.81, F1 macro = 0.735) The PCA data was classified with 73.5% accuracy using a StandardScalerWrapper, XGBoostClassifier model. The combined headline and article text features were classified with 82.24% accuracy using MaxAbsScaler, LightGBM.

	Headlines Only	Article Text Only	Headlines + Article Text
Keras model, raw data	0.7231	0.7545	0.8130
Keras model, PCA data	0.7197	0.7572	0.8051
Azure AutoML, raw data	0.7366	N/A	0.8224
Azure AutoML, PCA data	0.7352	N/A	N/A

It’s important to note that none of the models were trained on actual Facebook or Twitter posts (although links to many of the articles included can be found on Facebook and Twitter). While Facebook posts can be longer and often contain large sections of text resembling a full news article, tweets are limited to 280 characters. Because of this discrepancy, the developed models will likely perform differently on each site. Fake news on Twitter is typically presented as a simple headline and link to an article – some tweets will include a few key sentences from the article or serve as the anchor of a more detailed tweet thread. Because of this brevity, it is expected that model accuracy on Twitter will more closely resemble the 74% threshold found in the headline-only testing set. Facebook posts are typically more varied, ranging from an isolated headline to including an entire article. As a result, the accuracy of the models when applied to Facebook data may be as high as the 81.3% accuracy obtained by the headline + article testing.

An unfortunate conclusion to my literature review and my own model development is that no model which exclusively uses stylistic feature input was able to achieve a high enough accuracy to justify deployment as a fully-autonomous system in a non-experimental setting. Peak performance of 85-90% accuracy suggests that even the highest-performing models would be best implemented as part of a human-based AI system, with journalist fact-checkers acting as the final authority on decision-making. These moderately accurate models could prove very useful in speeding up the manual content selection and fact-checking process currently used by PolitiFact and other major fact-checking organizations. As stylistic models are improved and new techniques are developed, model accuracy will likely be able to transition into a sequential machine-human AI system, performing the majority of fact-checking independently with humans serving as sentinels (Saenz et al 2020). True 100% accuracy is required to trust the system in a fully-autonomous and unsupervised setting.

Like all data-driven models, the fake news detection models presented here are subject to the biases present in their training data. My model as well as the majority of models developed in previous research were trained on datasets compiled from existing fact-checking services, with PolitiFact being the most frequently used. While these services do strive to maintain integrity in the fact-checking process, there is likely to be small amount of residual bias, particularly in the verification of complex and nuanced political statements. A more apparent source of potential bias in the dataset comes with which claims are checked to begin with. PolitiFact relies on a team of journalists to scour the internet in search of check-worthy claims. They also maintain an email account where people can send links to potentially false claims. It is possible that the individual journalists, who tend to lean liberal, unintentionally introduce bias by choosing to fact-check conservative politicians and news outlets at a higher rate (Weaver el al 2018). If this were the case, the characteristic stylistic features observed in fake news may not be representative of the entire fake news media landscape; liberally-leaning fake news posts may exhibit a different set of characteristics that allow them to avoid detection by style-based classifiers.

Deployment

I believe this project would best be deployed using a primarily server-side model. The following is a general setup for deployment. The process begins when a given post reaches the threshold level of engagement to be fact-checked. To avoid inundating feeds with fact-checks, only posts that meet a predetermined engagement threshold (as measured by views, likes, or reposts) will be eligible. When a post crosses this threshold, the request to being the fact-checking process will be sent to the server. After the server receives the request, relevant NLP preprocessing steps are performed to extract the feature information required by the classification model. The data is first fed to a model to determine if the post was intended to be interpreted as news or not. The majority of posts will not be considered news and the process will exit. If the post does qualify as news, the server will run the fake news detection model and make a determination. At this point, the model may place an initial flag or warning label on the post while the human-side fact-checking process commences. After the post is confirmed as false, a more definitive label should be placed. The timing, prevalence, and actual content of the label is a significant factor in efficacy and should be informed by further research.

One reason for running the model on the server side as opposed to locally is that the model will not be run very often (relative to the total number of posts sent daily). Since only a subset of posts will be considered eligible for fact-checking, the longer network latency time is less of an issue. Additionally, the server request is only sent when a post reaches a certain threshold engagement, not at the instance of the user sending the post. This means that, from a user perspective, the longer network latency will largely go unnoticed. A second major factor in the decision to operate on servers is the high computational cost of running the model. While the model itself is not especially computationally expensive, extracting stylistic features from the post’s text is a slow process requiring a relatively high amount of computing power. Performing these computations locally could result in slower performance and ultimately affect user experience.

Following the deployment of the fake news detection model to a live social media environment, continued monitoring is necessary to ensure consistent performance. While stylistic concept drift in fake news text has not yet been observed, it remains a future possibility that the features initially responsible for a successful model no longer serve as reliable predictors of fake news (Minhas 2021). When automated fake news detection becomes a mainstream reality, publishers of fake news may be motivated to alter their signature brash and polarizing style in an effort to avoid detection. To keep concept drift under control, adversarial testing should be conducted throughout the development process, including after deployment. As part of the continuous monitoring process, a journalistic team will be required to confirm or correct the model’s classification. As continued adjustments are made and the model improves performance over time, this team will decrease in size and take on a less active role in the system. Other methods of model performance monitoring include the use of output checks and user feedback; however, it is likely that a majority of user feedback will be unreliable given the controversial status of fake news content. In addition to evaluating model performance, monitoring the efficacy of the system as a whole from a human-computer interaction perspective is very important. The studies referenced in the domain expertise section of this paper make clear that the visual presentation of the intervention (a warning label, flag, clarification, etc.) is as important to the success of the system as the model itself. Monitoring and continually testing the efficacy of the fake news labels, with adjustments made as necessary, is a vital factor in the success of this system.

Lastly, there are many ethical concerns to consider given this project’s potential for high profile in social and political life. Even given perfect performance, the polarizing views surrounding fake news in the political sphere make this project particularly prone to backlash. While aforementioned intended societal effect is to decrease the spread of harmful fake news, a strong enough backlash could end up having the opposite effect, causing social media users to reject the model’s findings en masse and become further entrenched in false beliefs. In a project with such a high potential to enact either positive or negative societal change, a robust and careful third-party auditing process is a necessity to ensure all ethical standards are met.

Summary

In recent years, fake news on social media has emerged as a serious threat to political discourse and the very existence of social media platforms. This paper examines strategies taken to detect fake news and proposes a novel automated system to slow the spread of fake news and reduce its harmful effects. The system proposed consists of a machine learning-based classifier in conjunction with a team of journalists, cognitive psychologists, and social media experts. Thus far, style-based fake news classifiers have exhibited moderate success, but continued improvement is necessary to progress closer to a fully-automated fake news detection system. The system proposed presents many technical, design-related, and ethical challenges, but I am confident that with continued research and commitment by social media companies, an effective and unproblematic system to detect and mitigate the spread of online fake news is feasible.

Bibliography

47 U.S.C. sec. 230, a Provision of the Communication Decency Act, 1996.

Ahmed H., Traore, I., Saad, S. (2018). Detecting Opinion Spams and Fake News Using Text Classification. Journal of Security and Privacy, 1(1).

Ahmed H., Traore, I., Saad, S. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. ISDDC 2017, 10619, 127-138.

Allyn, B. (2020). As Trump Targets Twitter's Legal Shield, Experts Have A Warning. npr.org

Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., & Huang, J. (2020). Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks. ArXiv:2001.06362 [Cs]. http://arxiv.org/abs/2001.06362

Bogaard, G., Meijer, E. H., Vrij, A., & Merckelbach, H. (2016). Scientific Content Analysis (SCAN) Cannot Distinguish Between Truthful and Fabricated Accounts of a Negative Event. Frontiers in Psychology, 7. https://doi.org/10.3389/fpsyg.2016.00243

Bond, G. D., Holman, R. D., Eggert, J.-A. L., Speller, L. F., Garcia, O. N., Mejia, S. C., Mcinnes, K. W., Ceniceros, E. C., & Rustige, R. (2017). 'Lyin' Ted', 'Crooked Hillary', and 'Deceptive Donald': Language of Lies in the 2016 US Presidential Debates: Language of lies in debates. Applied Cognitive Psychology, 31(6), 668–677. https://doi.org/10.1002/acp.3376

Bump, P. (2021). The Lie That Lingers: 3 in 10 Americans Falsely Believe the Election was Riddled with Fraud. Washington Post.

Chen, W., Lakshmanan, L., Castillo, C. (2013). Information and Influence Propagation in Social Networks. Synthesis Lectures on Data Management. Oct, 2013. https://doi.org/10.2200/S00527ED1V01Y201308DTM037

Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), 1–4. https://doi.org/10.1002/pra2.2015.145052010082

de Oliveira, N. R., Medeiros, D. S. V., & Mattos, D. M. F. (2020). A Sensitive Stylistic Approach to Identify Fake News on Social Networking. IEEE Signal Processing Letters, 27, 1250–1254. https://doi.org/10.1109/LSP.2020.3008087

ENISA. (2018). Strengthening Network and Information Security and Protecting Against Online Disinformation ("Fake News"). The EU Cybersecurity Agency.

Gharib, M. (2020). 'I Will Kill You': Health Care Workers Face Rising Attacks Amid COVID-19 Outbreak. NPR.

Giełczyk, A., Wawrzyniak, R., & Choraś, M. (2019). Evaluation of the Existing Tools for Fake News Detection. In K. Saeed, R. Chaki, & V. Janev (Eds.), Computer Information Systems and Industrial Management (Vol. 11703, pp. 144–151). Springer International Publishing. https://doi.org/10.1007/978-3-030-28957-7_13

Gstalter, M. (2020). Florida Poll Worker Details Alleged Harrassment, Assault From Trump Supporters: 'We were in a war zone'. TheHill.com.

Horne, B. D., Nørregaard, J., & Adali, S. (2020). Robust Fake News Detection Over Time and Attack. ACM Transactions on Intelligent Systems and Technology, 11(1), 1–23. https://doi.org/10.1145/3363818

Kavanagh, J., Marcellino, W., Blake, J., Smith, S., Davenport, S., & Tebeka, M. (2019). News in a Digital Age: Comparing the Presentation of News Information over Time and Across Media Platforms. RAND Corporation. https://doi.org/10.7249/RR2960

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. Journal of Experimental Psychology: Human Learning & Memory, 6(2), 107–118. https://doi.org/10.1037/0278-7393.6.2.107

Ksieniewicz, P., Zyblewski, P., Choras, M., Kozik, R., Gielczyk, A., & Wozniak, M. (2020). Fake News Detection from Data Streams. 2020 International Joint Conference on Neural Networks (IJCNN), 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207498

Langeveld, M. (2009). The Content Cascade: How content will flow in digital news enterprises. Nieman Foundation for Journalism. https://www.niemanlab.org/2009/04/managing-the-content-cascade/

Lawrence, N.D. (2017). Data Readiness Levels. ArXiv:1705.02245

Lawrence, N. D. (2019). Data Science and Digital Systems: The 3Ds of Machine Learning Systems Design. ArXiv:1903.11241

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lee, M.Y.H., Brown, E. (2020). Election Officials Warn Trump's Escalating Attacks On Voting Are Putting Their Staffs At Risk. Washington Post.

Levin, B. (2021). Report to the Nation: Anti-Asian Prejudice & Hate Crime. Center for the Study of Hate & Extremism CSUSB. 2020-21 First Quarter Comparison Data.

Matsa, K.E., and Shearer, E. (2018) News Use Across Social Media Platforms 2018. Pew Research Center.

Miller, J. (2020). Do COVID-19 Conspiracy Theory Beliefs Form a Monological Belief System? Canadian Journal of Political Science, 53(2), 319-326. doi:10.1017/S0008423920000517.

Minhas, K. (2021). An Assessment of Concept Drift in Style-Based Fake News Classifiers.

Moravec, P., Minas, R., & Dennis, A. R. (2018). Fake News on Social Media: People Believe What They Want to Believe When it Makes No Sense at All. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3269541

Mourão, R. R., & Robertson, C. T. (2019). Fake News as Discursive Integration: An Analysis of Sites That Publish False, Misleading, Hyperpartisan and Sensational Information. Journalism Studies, 20(14), 2077–2095. https://doi.org/10.1080/1461670X.2019.1566871

Mustafaraj, E., & Metaxas, P. T. (2017). The Fake News Spreading Plague: Was it Preventable? ArXiv:1703.06988 [Cs]. http://arxiv.org/abs/1703.06988

Oshikawa, R., Qian, J., & Wang, W. Y. (2020). A Survey on Natural Language Processing for Fake News Detection. ArXiv:1811.00770 [Cs]. http://arxiv.org/abs/1811.00770

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. Management Science, 66(11), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

PolitiFact. Truth-o-meter. https://www.politifact.com/truth-o-meter/

Rapoza, K. (2017). Can 'Fake News' Impact the Stock Market? Forbes

Reis, J. C. S., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Explainable Machine Learning for Fake News Detection. Proceedings of the 10th ACM Conference on Web Science - WebSci '19, 17–26. https://doi.org/10.1145/3292522.3326027

Reis, J. C. S., Correia, A., Murai, F., Veloso, A., Benevenuto, F., & Cambria, E. (2019). Supervised Learning for Fake News Detection. IEEE Intelligent Systems, 34(2), 76–81. https://doi.org/10.1109/MIS.2019.2899143

Rogers, R. (2020). Research note: The scale of Facebook's problem depends upon how 'fake news' is classified. Harvard Kennedy School (HKS) Misinformation Review. https://doi.org/10.37016/mr-2020-43

Rollwage, M., & Fleming, S. M. (2021). Confirmation bias is adaptive when coupled with efficient metacognition. Philosophical Transactions of the Royal Society B: Biological Sciences, 376(1822), 20200131. https://doi.org/10.1098/rstb.2020.0131

Romer, D., Jamieson, K.H. (2020). Conspiracy Theories as Barriers to Controlling the Spread of Covid-19 in the U.S. Social Science & Medicine, 263. https://doi.org/10.1016/j.socscimed.2020.113356

Saenz, M.J., Revilla, E., Símon, C. (2020). Designing AI Systems With Human-Machine Teams. MIT Sloan Management Review.

Said, C. (2020). California Officers Facing Protests, Even Death Threats, Over Coronavirus Orders. San Francisco Chronicle.

Seo, H., Xiong, A., & Lee, D. (2019). Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation. Proceedings of the 10th ACM Conference on Web Science - WebSci '19, 265–274. https://doi.org/10.1145/3292522.3326012

Shu, K., & Liu, H. (2019). Detecting Fake News on Social Media. Synthesis Lectures on Data Mining and Knowledge Discovery, 11(3), 1–129. https://doi.org/10.2200/S00926ED1V01Y201906DMK018

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. Big Data, 8(3), 171–188. https://doi.org/10.1089/big.2020.0062

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36. https://doi.org/10.1145/3137597.3137600

Shu, K., Wang, S., & Liu, H. (2019). Beyond News Contents: The Role of Social Context for Fake News Detection. Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 312–320. https://doi.org/10.1145/3289600.3290994

Shu, K., Zhou, X., Wang, S., Zafarani, R., & Liu, H. (2019). The role of user profiles for fake news detection. Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 436–439. https://doi.org/10.1145/3341161.3342927

Silverman

Su, Q., Wan, M., Liu, X., & Huang, C.-R. (2020). Motivations, Methods and Metrics of Misinformation Detection: An NLP Perspective. Natural Language Processing Research. https://doi.org/10.2991/nlpr.d.200522.001

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some Like it Hoax: Automated Fake News Detection in Social Networks. ArXiv:1704.07506 [Cs]. http://arxiv.org/abs/1704.07506

Taylor, L. (2020). Covid-19 Misinformation Sparks Threats and Violence Against Doctors in Latin America. BMJ 2020(370). https://doi.org/10.1136/bmj.m3088

Torabi Asr, F., & Taboada, M. (2019). Big Data and quality data for fake news and misinformation detection. Big Data & Society, 6(1), 205395171984331. https://doi.org/10.1177/2053951719843310

Traylor, T., Straub, J., Gurmeet, & Snell, N. (2019). Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator. 2019 IEEE 13th International Conference on Semantic Computing (ICSC), 445–449. https://doi.org/10.1109/ICOSC.2019.8665593

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wang, R., He, Y., Xu, J., & Zhang, H. (2020). Fake news or bad news? Toward an emotion-driven cognitive dissonance model of misinformation diffusion. Asian Journal of Communication, 30(5), 317–342. https://doi.org/10.1080/01292986.2020.1811737

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 849–857. https://doi.org/10.1145/3219819.3219903

Wang, W.Y. (2017). ‘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, 2, pp. 422–426.

Weaver, D.H., Willnat, L., Wilhoit, G.C. (2018). The American Journalist in the Digital Age: Another Look at U.S. News People. Journalism and Mass Communication Quarterly. 96(1), 101-130. https://doi.org/10.1177/1077699018778242.

Wineburg, S., McGrew, S., Breakstone, J., & Ortega, T. (2016). Evaluating Information: The Cornerstone of Civic Online Reasoning. Stanford Digital Repository. http://purl.stanford.edu/fv751yt5934

Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., & Liu, H. (2019). Unsupervised Fake News Detection on Social Media: A Generative Approach. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 5644–5651. https://doi.org/10.1609/aaai.v33i01.33015644

Zengerle, P. (2016). Clinton Calls ‘Fake News’ a Threat to U.S. Democracy. Reuters.com.

Zhang, S., & Kejriwal, M. (2019). Concept drift in bias and sensationalism detection: An experimental study. Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 601–604. https://doi.org/10.1145/3341161.3343690

Zhou, X., & Zafarani, R. (2019). Network-based Fake News Detection: A Pattern-driven Approach. ACM SIGKDD Explorations Newsletter, 21(2), 48–60. https://doi.org/10.1145/3373464.3373473

Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Computing Surveys, 53(5), 1–40. https://doi.org/10.1145/3395046

Zhou, Z., Guan, H., Bhat, M. M., & Hsu, J. (2019). Fake News Detection via NLP is Vulnerable to Adversarial Attacks. Proceedings of the 11th International Conference on Agents and Artificial Intelligence, 794–800. https://doi.org/10.5220/0007566307940800

Zuckerberg, M. (2020). Facebook, Social Media Privacy, and the Use and Abuse of Data. Senate Committee on the Judiciary, Senate Committee on Science, and Transportation.