CoVerifi: A COVID-19 News Verification System

Nikhil Kolluri[a], Dhiraj Murthy[b,1]

[a] Department of Electrical and Computer Engineering, University of Texas, Austin, TX 78712, USA

[b] School of Journalism and Media, Moody College of Communication and Department of Sociology, University of Texas at Austin, Austin, TX 78712, USA

Abstract

There is an abundance of misinformation, disinformation, and “fake news” related to COVID-19, leading the director-general of the World Health Organization to term this an ‘infodemic’. Given the high volume of COVID-19 content on the Internet, many find it difficult to evaluate veracity. Vulnerable and marginalized groups are being misinformed and subject to high levels of stress. Riots and panic buying have also taken place due to “fake news”. However, individual research-led websites can make a major difference in terms of providing accurate information. For example, the Johns Hopkins Coronavirus Resource Center website has over 81 million entries linked to it on Google. With the outbreak of COVID-19 and the knowledge that deceptive news has the potential to measurably affect the beliefs of the public, new strategies are needed to prevent the spread of misinformation. This study seeks to make a timely intervention to the information landscape through a COVID-19 “fake news”, misinformation, and disinformation website. In this article, we introduce CoVerifi, a web application which combines both the power of machine learning and the power of human feedback to assess the credibility of news. By allowing users the ability to “vote” on news content, the CoVerifi platform will allow us to release labelled data as open source, which will enable further research on preventing the spread of COVID-19-related misinformation. We discuss the development of CoVerifi and the potential utility of deploying the system at scale for combating the COVID-19 “infodemic”.

Keywords: Infodemic, Misinformation, Media Diet, Machine Learning, COVID-19

1. Introduction

Coronavirus (COVID-19), declared a Public Health Emergency of International Concern (PHEIC), is a virus which originated in Wuhan, China in December 2019 [1]. As of November 22, 2020, COVID-19 has spread to 220 countries, areas, or territories, infected over 57.8 million people, and killed over 1.3 million people [2]. In February 2020, World Health Organization (WHO) Director-General Tedros Adhanom Ghebreyesus said “We’re not just fighting an epidemic; we’re fighting an infodemic. Fake news spreads faster and more easily than this virus, and is just as dangerous.” [1]. Sylvie Briand, architect of the WHO’s strategy to counter the infodemic risk argues that “with social media [...] this phenomenon is amplified, it goes faster and further, like the viruses that travel with people and go faster and further” [1]. Previous work examining COVID-19 tweets found that twice as much false information as evidence-based information was tweeted, though this trend did not apply to retweets [3]. Among the Twitter

[1] Corresponding author. Email address: Dhiraj.Murthy@austin.utexas.edu (Dhiraj Murthy)

Version of record to be published in: *Online Social Networks and Media*

posts regarding COVID-19 in their sample, most posts (59%) rated false by their fact-checkers remained up at the time of publishing their article [4].

This infodemic has proven its ability to accelerate the epidemic process, increase violence against certain groups, and cause bodily harm. COVID-19 misinformation can directly threaten lives. There are harmful "cures" being suggested such as drinking fish tank additives, bleach, or cow urine [5]. Furthermore, there exists a threat of COVID-19 misinformation increasing resistance to a vaccine. There is already a growing anti-vaccination community related to COVID-19, which according to some, is better positioned for growth than the pro-vaccination community [5]. Additionally, false rumors that people with dark skin may be immune to COVID-19 have been spreading on social media since late January 2020, and have potentially contributed to the over-representation of some minorities as victims. In the US, as of early April 2020, approximately 70% of fatalities in Chicago and Louisiana were African Americans, who only make up roughly 30% of the population [5] [6] [7]. APM Research Lab found that Black, Indigenous, and Latino Americans all had a COVID-19 death rate of triple or more White Americans [8]. Moreover, some malicious COVID-19 narratives have been linked to offline anti-Asian violence [5] [9] [10].

To help address the continuing and major impacts associated with COVID-19-related misinformation, this article introduces CoVerifi, our web application to assess the credibility of COVID-19 news. CoVerifi retrieves a selection of news articles, tweets, and Reddit posts and displays credibility ratings produced by machine learning models and a credibility rating obtained from the "votes" of other CoVerifi users, and for tweets, the Botometer API's [11] [12] bot score for the tweet poster. As the platform is used, we plan to release the labelled data we collect from these user votes as an open source dataset, which will enable further research on preventing the spread of COVID-19-related misinformation. As CoVerifi is open source, we also enable important future research, including the expansion of CoVerifi to other news-related platforms, machine learning models, and data collection opportunities. We discuss the development of CoVerifi and the potential utility of deploying the system at scale for combating the COVID-19 infodemic.

1.1. Fake news, Misinformation, and Disinformation

Lazer et al. [13] define "fake news" "to be fabricated information that mimics news media content in form but not in organizational process or intent". As Tandoc et al. [14] highlight, the term remains contentious, yet important to engage with; hence, our inclusion of quotation marks. Gelbert [15] acknowledges the term "fake news" as new and rapidly evolving but argues in favor of only using it to refer to misleading claims which are misleading by design. To best address the scope of our research, our use of "fake news" includes inaccurate news, low-quality news, and imposter news. Moreover as Lazer et al. [13] observe 'fake news overlaps with other information disorders, such as misinformation (false or misleading information) and disinformation (false information that is purposely spread to deceive people)'. Tandoc et al. [14] add that "fake news" also encompasses 'viral posts based on fictitious accounts made to look like news reports'.

These types of "fake news" can be split into two categories: "fake news" with an implicit understanding by the reader that the content is false (such as parody and satire), and "fake news" where readers are unaware that the information is false. A recent study found that 38% of COVID-19-related misinformation in their sample was completely fabricated, 59% of the misinformation involved reconfiguration, and only 3% was satire/parody [4]. This suggests that amidst COVID-19, "fake news" which the reader is unaware is fake is present at alarming rates.

This category can be further broken into two groups: misinformation, which refers to the “inadvertent sharing of false information” [16] [17] and disinformation, which refers to “the deliberate creation and sharing of information known to be false” [16] [17]. The spread of misleading information on the web and social media “poses a major risk to society” [18] and “is overloading the exchange of ideas upon which democracies depend” [19]. With the presence of algorithms which personalize online experiences and hinder exposure to ideologically diverse sources of information, some argue that echo chambers emerge which make it harder to encounter ideologically diverse types of information [19] [20]. Since social media and digital platforms are capable of substantially fragmenting the public’s opinions and decreasing challenges against untrue information, we must very seriously consider strategies to combat misinformation which traverses these platforms. Moreover, for misinformation on social media and digital platforms, factual corrections are often ineffective, slow, and rarely reaching the people originally influenced by the misinformation [19]. Since social media and digital platforms are characterized by providing a massive quantity of information without the ability to provide factual corrections on content that has already been consumed, analysis of the way consumers address future media content, the types of media content provided to users, and the potential to foster organic responses to misinformation is increasingly important. Disruptions in the information landscape caused by COVID-19 warrant even greater consideration of these lines of research, since the veracity of content that individuals are consuming has measurable health impacts [21] [5].

1.2. Pre-COVID-19 Misinformation Research

Prior to the outbreak of COVID-19, there were several attempts at addressing the more general problem of combating “fake news” and misinformation via a semi-automated or fully-automated approach. Past research reached a conclusion that social media is systematically exploited to manipulate and alter the public opinion [22]. Since these attacks are often orchestrated using bots, previous work has used machine learning to separate humans from bots, which can be used to determine if a social media account is part of a nefarious campaign [22] [23]. Furthermore, there are several works surrounding computer-aided strategies to combat “fake news”, including stance-detection machine learning models, neural “fake news” detection models, claim identification pipelines, and “fake news” datasets. Collectively, these address “fake news” from news articles, Reddit posts, Facebook posts, and Twitter tweets. The inclusion of information from a variety of platforms is important because it parallels the multiplatform nature of the media diet of the average person. The notion that media habits are best characterized by diverse media consumption patterns has been discussed for decades, with early work involving ‘time diaries’ which qualitatively document media habits [24]. Today, people continue to consume information from diverse sources and prioritize their time on content from certain sources and even on specific topics. In the case of Twitter, users tend to consume information primarily on one or two specific topics of their interest, but the Twitter recommendation system mitigates imbalances in users’ consumed diets [25]. Users therefore can be presented with information that is unbalanced in terms of coverage of news stories and different from what other users are presented with [25]. Since it is typical for news consumers to draw from a range of sources, an approach which leverages a variety of information sources is crucial.

Social media has been shown to play a very important role in the media diet of digital native voters. Research has indicated that a digital media environment may socialize young voters into polarized information environments which in turn increases their involvement in elections [26]. Common in misinformation research prior to the outbreak of COVID-19 was the use of machine learning, a subset of artificial intelligence “that involves building and adapting models, which allow programs to ‘learn’ through experience” [27]. A ‘model’ represents something which takes in data as an input, performs some computation on the data, and produces some information as an output. Furthermore, when discussing machine learning models for news verification purposes, the models can often be described as performing natural language processing (NLP), a type of artificial intelligence tasked with comprehending, deciphering, and even reproducing human languages.

1.2.1. Existing Models and Approaches

There has been substantial work in leveraging machine learning and artificial intelligence to differentiate between fake and real information. Based on the types of machine learning models and web tools which have already been developed, it appears difficult to create a robust, entirely feature-based NLP model which includes no external information. Even seemingly performant natural language processing models have shown significantly reduced accuracy when presented with reconfigured news (changing small amounts of information to make the information false) as part of an adversarial attack [28]. Therefore, much of the prior work seems to be on seemingly peripheral tasks, such as stance detection, neural “fake news” detection, bot detection, and multi-step approaches involving the inclusion of external information.

An important step in developing the notion that “fake news” detection is not best addressed as a singular, isolated machine learning model was the “fake news” Challenge Stance Detection Task (FNC1). Notably, the competition focused on the task of stance detection (i.e., evaluating whether the headline agrees with the claim) rather than the task of labelling a claim as true or false [29]. The FNC-1 creators found that “truth labeling” is very difficult in practice and preferred a reliable semi-automated tool over a fully-automated system, which they felt would inevitably fall far short of 100% accuracy [29]. The results of FNC-1 encourage future work toward tools which aid journalists and fact checkers, both because fact checkers and journalists have acknowledged semi-automated tools as valuable and because machine learning in a vacuum may not be able solve the truth labelling problem.

With the presence of text generation models such as Google’s BERT [30] which do a good job at mimicking real speech patterns, neural “fake news” generated by robots became reality. Since neural “fake news” algorithms cause the resulting generated texts to have some similar traits, achieving high accuracy with a neural “fake news” detection model is possible. As a response, several models have emerged to detect neural “fake news”. These include Grover and GPT-2. The former is a “fake news” generator which can also spot “fake news” generated by other AI models. In a setting with a limited access to neural “fake news” articles, Grover obtained over 92% accuracy at differentiating between human-written and machine-written news [31]. GPT-2, a successor to GPT, was trained to predict the next word in internet text, which allows it to generate synthetic text [32]. OpenAI also released a GPT-2 output dataset, which was used to train a corresponding fake text detector model capable of labelling text as “real” or “fake” with a confidence percentage [33]. An important note is that this fake text detector model will likely perform best on text generated by GPT-2, though our intuition is that it may help identify fake text generated by other models which use similar text-generation algorithms. The GPT-2 output detector should be seen as a tool for predicting whether content was generated by a machine or by a human. It is not capable of directly predicting veracity. For a more in-depth discussion of this important distinction, see Section 2.2.3.

Other work tangential to neural “fake news” detection includes bot detection. The research surrounding bot detection is highly relevant to “fake news” since it can suggest that “fake news” is often propagated as a result of orchestrated, malicious campaigns. BotOrNot, now renamed Botometer, was a system to evaluate social bots which has served more than one million requests via their website. While BotOrNot provides effective functionality for samples up to thousands of accounts, it cannot scale much more extensively than that due to its reliance on the rate-limited Twitter API [22]. Other work utilized a mixture of machine learning techniques and cognitive heuristics for bot detection [22]. Ferrara et al. [22] found that bots that existed during the 2016 U.S. Presidential election campaign to support alt-right narratives went dark after November 8, 2016, and were used again in the days prior to the 2017 French presidential election. Moreover, it is reasonable to conclude that their content would be overrepresented in relation to the amount of users orchestrating the campaign, given that bots can produce far more content than humans in short timescales. Therefore, addressing these intentional, bot-driven, malicious campaigns remains important to the literature due to (1) the relative ease of identifying neural “fake news” and bot-created news compared to identifying a diverse range of “fake news” types and (2) the ability to identify large volumes of “fake news” at once if the presence of a bot is detected. Given this discussion of the utility of models identifying bot-generated news in combating misinformation at large, we chose to use a form of neural “fake news” detection (a text classification model trained on the outputs of GPT-2) in our CoVerifi platform. For tweets, we also include the bot score provided by Botometer’s API for the poster’s account.

However, it should be noted that automated approaches to assess the validity of a piece of news content have had some success, though currently reported accuracy has not yet reached acceptable levels. This research included the Fact Extraction and VERification (FEVER) Shared Task, which challenged participants to classify whether human-written factoid claims could be supported or refuted by using evidence retrieved from Wikipedia. The best performing system achieved a FEVER score of 64.21% [34]. This line of research demonstrates that while creating a fully automated machine-learning based approach for assessing the validity of a piece of news may indeed be possible, it seems to be a difficult, computationally intensive process with potentially marginal gains. This information strengthens the claim that pursuing tools that make fact checking easier and more effective could potentially be a more rewarding research area than an entirely automated approach.

1.2.2. Existing Web Tools

Toward the end of creating tools that make fact checking easier rather than producing a single, insular, machine-learning based approach, there are several web-based tools which have made substantial progress. Among these are ClaimBuster, Google Fact Check, and GLTR. ClaimBuster offers a near-complete fact-checking system, whereas Google Fact Check allows users to check specific claims and GLTR allows a visualization of the likelihood that text is machine-generated. Each provides specific and unique value, all advancing the goal of mitigating the harm caused by the spread of misinformation.

ClaimBuster [35] monitors live discourses (interviews, speeches, and debates), social media, and news to identify factual claims, detect matches with a professionally-verified repository of fact-checks, and instantly deliver the result to the audience [36]. For new, unchecked claims, ClaimBuster translates them into queries against knowledge databases and reports the result [36]. If humans must be brought into the loop for claims, it provides tools to help lay people to understand and vet claims [36]. This decision to provide tools rather than a classification for cases in which humans must be brought into the loop reveals that the authors were aware that an insulated, entirely automated approach may not be sufficient. While the decision-making process behind ClaimBuster is a great step in a positive direction, the system has limitations. It appears to focus on U.S. Presidential Election information, the ability to match claims against existing knowledge bases, and tools to check tweets. Therefore, ClaimBuster has limitations in terms of being able to address a wide range of types, formats, and quantities of misinformation, especially since more subtle forms of misinformation may not be possible to query against existing knowledge bases, and may occur on platforms which the authors are not aware of. Moreover, while a targeted approach is powerful and may work for many types of misinformation, it does not provide a solution which can change and expand at the same rate as the quantity and type of misinformation expands. Ultimately, such approaches could be complemented by less-perfect, but more-scalable approaches which can grow at the same rate as misinformation grows.

Other web tools which address similar goals include the Google Fact Check Explorer and GLTR. The former is a resource which allows searching for a claim and receiving information from fact checking sites which have rated the claim as likely true or likely false. For example, searching “inhale steam to kill coronavirus” will return several results from similar claims which were fact checked (in this case, rated “False”) by different sources [37]. There is an associated Google FactCheck Claim Search API which can be used to query the same set of fact check results as the Fact Check Explorer. While the Google Fact Check Explorer is useful, it can only provide help when there are specific claims which need to be checked. It is not designed to evaluate a verbatim new article, tweet, or post. Thus, Google Fact Check Explorer seems to be best used as a tool for types of “fake news” which involve specific, previously manually-checked claims, but is not well-suited to evaluate large quantities of misinformation in varying formats. GLTR, a tool to detect automatically 5 generated text [38], has access to the GPT-2 and BERT language models and, given a textual input, can analyze what the models would have predicted as the next word for any position [38]. Using this knowledge, GLTR can help visualize the likelihood that a text passage is fake by using different colors to show which words were within the top words that these language models would have predicted [39] [38]. GLTR is similar to Google Fact Check Explorer in that it also serves a useful, but specific and limited function. While it could help a user to determine if a body of text is machine generated, it seems to require manually submitting information to the tool every time it is used, which again limits its scalability.

Additional projects working toward the goal of partially-automated or fully-automated misinformation detection include those partnered with SOMA [40] [41] [42]. The European Union’s SOMA project contains an Observatory, which aims to support experts in their work against disinformation by providing them with cyberinfrastructure and a network of people working on misinformation detection [40] [41] [42]. Members of the Observatory have access to existing verification platforms as well as new tools and algorithms [40]. All members of the SOMA project are given access to EUNOMIA, a platform for analyzing the source, modification history, and trustworthiness of a piece of content which includes blockchain-based infrastructure, a digital companion which uses AI to analyze content and context, and the ability to vote on the trustworthiness of social media posts [40] [43]. The platform requires the consent of the user to have the posts in their social media analyzed for trustworthiness [44] [43]. SOMA members are also given access to SocialTruth, which provides individuals with access to “fake news” detection based on AI technology and content verification trust and integrity based on blockchain technology [45] [46]. SocialTruth integrates its content verification with various platforms such as web search, journalist tools, content development, and a browser addon [45] [46]. WeVerify aims to address content verification challenges through participatory verification, open source algorithms, human-in-the-loop machine learning, and visualizations [47] [48]. The project is designed for collaborative, decentralized content verification, tracking, and debunking [47] [48] . WeVerify has specific elements designed for journalists, such as tools for detecting the spread of misinformation. Their inVID plugin, for example, is particularly useful for fact checking content, including video [49]. The Provenance project seeks to develop an intermediary-free solution for digital content verification [50]. The project claims that its solutions will make it easier for consumers to evaluate online information by providing a graphical guide that will clarify the source and history of a piece of content [50]. Moreover, work by the Provenance team found that "countermeasures which encourage citizens to reflect on the information they consume and choose to share is likely to be more effective than authoritative corrections" [51].

1.2.3. Existing Datasets

For the goal of enabling research on strategies to combat misinformation, there exists several potentially useful datasets. The LIAR dataset consists of 12.8K manually labeled short statements from politifact.com ranked as barely true, false, half true, mostly true, or pants on fire [52]. Other well known datasets includes the ISOT dataset, which consists of 21,417 real news articles and 23,481 "fake news" articles [53] [54] and 1000 news articles, evenly split between fake and legitimate news [55]. The presence of multiple datasets with misinformation content in multiple formats is useful, since the types of "fake news" experienced amidst the COVID-19 infodemic are broad and span multiple formats. One Twitter-specific dataset is CREDBANK, a crowdsourced dataset of accuracy assessments for events in Twitter, and another is PHEME, a dataset of potential rumors in Twitter and journalistic assessments of their accuracy [56]. Furthermore, BuzzFeed created a "fake news" dataset consisting of Facebook news, but their dataset has been used by Buntain et al. to extract parallel "fake news" data from Twitter [39] [40]. Interestingly, one study found that models trained against the crowdsourced workers dataset (CREDBANK) outperformed models trained against the journalists' assessment dataset (PHEME) when tested on Twitter data sourced from the BuzzFeed "fake news" dataset [39]. This is significant as it 6 indicates that crowd-sourced data may be useful. Specifically, if crowd-sourced information is effective, platforms leveraging this approach could be powerful tools for rapidly developing user interfaces that allow crowd-sourced information collection across multiple platforms. Furthermore, prior analysis has found that certain datasets overrepresented some topics and underrepresented others [57]. When a model is trained on an unbalanced dataset and exposed to a new type of data, it is liable to arbitrarily place the new data in the wrong class [57]. Moreover, there have been calls to action to create a greater volume of reliably labelled "fake news" articles [57].

1.3. COVID-19 Misinformation Research

After the outbreak of COVID-19, rich literature has emerged surrounding the amount of misinformation, the type of misinformation, and approaches to combat misinformation. This work has examined global trends on Twitter by country, analyzing tweet volume according to specific themes in coronavirus-related queries, posts related to specific myths surrounding the virus, and the number of tweets containing items deemed myths [58]. Other work attempts to find warning signs that a country will experience an infodemic [58]. Furthermore, studies have shown that hateful content is rapidly evolving and becoming increasingly coherent as time continues [58].

1.3.1. Results of Research Characterizing the Infodemic

A significant finding supporting the existence of an infodemic surrounding COVID-19 is that fact checkers are overburdened. One study found that after the outbreak of COVID-19, the number of English-language fact checks increased by 900% from January to March [4]. Despite this increase in the number of fact checks, fact checkers have limited resources and cannot check all problematic content [4]. In addition to research characterizing the types and quantities of information shared as news, there has been similar research characterizing traits of COVID-19 information shared on social media platforms, such as Twitter. One finding is that some keywords are correlated with misinformation. A study of 673 tweets over 14 hashtags and keywords related to the COVID-19 epidemic found that 24.8% of the sample included misinformation, 17.4% included unverifiable information, and tweets from unverified Twitter accounts contained more misinformation [59]. Other work indicates that machine learning can be used to get information about the key phrases used by people discussing the pandemic, as well as the emotional sentiment among phrases in these groups [60]. The knowledge that machine learning can be used for sentiment analysis of tweets is significant since it would potentially allow research discussing the relationship between hateful COVID-19 misinformation and the sentiment of the content.

1.3.2. Research on Combating COVID-19 Misinformation

At the heart of the explosion of misinformation present in the COVID-19 infodemic is the question of what makes people share misinformation. A study found that people shared false claims related to COVID-19 partly because they didn’t think sufficiently about whether or not the content was accurate before deciding what to share [61]. When participants were primed to think about accuracy at the beginning of a study, their level of truth discernment in whether they intended to share the COVID-19- related articles was more than doubled [61]. The conclusion was that priming individuals to think that accuracy is important to consider can significantly affect their truth discernment [61]. This work indicates that “truth nudging”, or priming individuals to think about accuracy of news content, could serve as a highly effective form of misinformation treatment while requiring virtually no computational power and potentially outperforming machines for certain types of misinformation. It also means that research on automated approaches should be carefully examined to ensure that the automated approach does not give a false illusion of certainty. If an automated approach claims that a piece of “fake news” content is real, it may make the user less likely to critically examine it, which in turn could decrease the user’s truth discernment.

Other work explores the impact of providing a news feed that has been vetted for factuality to users. WashKaro uses NLP approaches, machine learning, and m-Health to provide authentic sources of information with daily news in Hindi and English, along with performing other functions such as contact tracing [62]. It provides authentic news by checking for similarity between new news articles and existing news articles in their dataset that were clustered according to which WHO guidelines they are similar to; when new guidelines are added, news articles with the 10 highest similarity ratings are provided to the user [62]. WashKaro provides information related to health guidelines, rather than COVID-19 information holistically. For example, their application does not seem to apply to racially-targeted hateful misinformation.

Other work examined malicious COVID-19 content and found that hateful COVID-19 content mobilizes and accumulates on platforms which allow specific community features (such as Facebook Pages / VKontakte groups), but then makes its way back to the mainstream [9]. The study also found that real, racially motivated violence occurred after the outbreak of COVID-19, suggesting potential implications of these hateful communities [9]. While malicious activity can appear isolated and largely eradicated on a given platform, it has likely just moved to another platform [9]. One possible research area this line of work opens up is on the effectiveness of utilizing the “truth nudging” strategy on a given news or media platform to let users know that when they leave one platform to visit another platform, there is a risk of encountering hateful or malicious content. Another potential research area, given the presence of hateful COVID-19 information, is determining the effectiveness of implementing sentiment detection (determining the predominant emotion in a text) using natural language processing to identify potentially hateful content. Finally, since malicious COVID-19 information is heavily decentralized and regulations on a single platform may simply push the information to other platforms, research questions evaluating whether an open-source, multi-platform misinformation detection tool could be useful become important. Perhaps one response to the proliferation of a variety of deregulated, open-source media platforms is a symmetric open-source, multi-platform misinformation detection tool which could be altered to allow usage on a wide variety of platforms.

Other approaches acknowledge that the massive volume of new online material surrounding COVID-19 makes manual analysis non-viable, opening the door to automated machine learning approaches to combat misinformation through counter-messaging [5]. It is not yet known whether overt targeted ads presenting counter-information would be as effective as a more holistic truth-nudging approach, which has been shown to double truth discernment in social media content sharing decisions [61]. Ultimately, it is clear that it is virtually impossible to address all of "fake news" in the form of a singular, generalizable solution. However, targeted approaches, like WashKaro's use of the WHO's guidelines to decide which news to share, have downsides in their ability to address new types of information. One way to address this is creating a tool which could be used to do a good enough job at detecting "fake news" on a wide variety of platforms, performs truth nudging to encourage users to critically reflect on the veracity of what they have read, and warns users of the dangers of leaving one platform for another.

1.4. Additional Noteworthy COVID-19 Misinformation Research

There have been several studies that focused on the role played by Facebook in online discussions surrounding COVID-19. Research studying rumors claiming that the rollout of 5G technology was related to the outbreak of the COVID-19 pandemic found that conspiracy theories can start as ideas that move from fringe beliefs to entering mainstream discourse [63]. Therefore, understanding typical processes of idea spread is an important method for gaining a clearer picture of key points at which dissemination may be slowed or halted [63]. Other work which examined the role of Facebook advertisements in facilitating coronavirus-related conversation found instances of possible misinformation ranging from bioweapons conspiracy theories to unverifiable claims by politicians [64]. Non-English-language work using over 1.5 million Italian-language posts on Facebook related to COVID-19 found that sources of 'supposedly' reliable information experienced higher engagement compared to websites sharing unreliable content with a "small-world effect" observed in the sharing of URLs [65]. Moreover, users who navigate a limited set of pages/groups can be exposed to a wide range of content, ranging from extreme propaganda to verified information [65]. A study of alternative news media content on Facebook used computational content analysis to evaluate the validity of the claim that alternative news media outlets spread societal confusion and potentially dangerous 'fake news' [66]. The authors found that while alternative news media outlets do not tend to spread obvious lies, they do predominantly share critical messages, including anti-establishment views (which oppose mainstream news media and the political establishment) that can contribute to worldviews based on mistrust [66].

The literature studying the role of Twitter in online COVID-19 conversations is considerable. A study of two competing COVID-19 misinformation communities - misinformed users (who are actively posting misinformation) and informed users (who are spreading true information) - concluded that COVID-19 misinformed communities are denser and more organized than informed communities [67]. A significant volume of the misinformation they studied likely originated in disinformation campaigns, a large majority of misinformed users may be anti-vaxxers, and informed users tend to use more narratives than misinformed users [67]. A study of 43.3 million English-language tweets related to COVID-19 found evidence of the presence of bots in the COVID-19 discussion on Twitter [68]. High bot score accounts were found to use COVID-19-related content and hashtags to promote visibility of ideological hashtags that are typically associated with the alt-right in the United States and human users, on the other hand, are predominantly concerned with public health and welfare [68]. Research exploring high and low quality URLs shared on Twitter in the context of COVID-19 found that more tweets contained URLs from low quality misinformation websites compared to high quality health information websites, but both are present at a much lower rate compared to news sources [69]. While some high and low quality sites are connected, connections to and from news sources are more common [69]. The authors' findings suggest that despite low quality URLs not being extensively shared in the COVID-19 Twitter conversation, there is "a well connected community of low quality COVID-19 related information" which has connections to both health and news sources [69]. Using a dataset of 67 million tweets from 12 million users, other work found that the majority of influential tweets were posted by news media, government officials, and individual news reporters, but the most influential tweets were posted by average, everyday users [70]. They observed that average users were also more likely to spread noncredible tweets, but that "many of these regular users appear to be bots" [70].

2. Evaluation and Development

2.1 Development Considerations

We sought to rapidly and affordably develop a web tool leveraging natural language processing and highly accessible human feedback to provide an experience which has an accessible, easy-to-use experience and draws from a variety of sources to match a realistic COVID-19-related media diet. The project's development occurred from May-July 2020. Our tool was designed to be able to aid in contributing to the body of labelled data surrounding "fake news" and especially COVID-19 "fake news". These key decision-making factors are described below:

Combining NLP and Human Feedback. As demonstrated in the analysis on prior models, it is difficult to create a completely accurate, robust, and fully-automated machine learning approach to "fake news" detection. Thus, while we decided to use machine learning, we also felt that the combination of machine learning with human feedback was very valuable, since it allows for misclassifications by the ML model to be detected by humans and for "fake news" which fools humans to be detected by ML models. This allows us to leverage ML without creating an illusion of absoluteness behind the results and enables our service to remain useful even for ML model "edge cases," such as mostly true news with a few key words replaced to make the claims false.

Accessible and Realistic News Feed. Prior approaches are effective at helping users identify if a claim or body of text are rated true or false by machine learning models, but are often either for specific types of textual input or would be cumbersome to include in one's typical news consumption process. Therefore, through attempting to create an accessible news feed which is representative of the typical "media diet" of an average person, we aim to increase the likelihood that our fact checking tools will actually be used. Creating a dynamically generated feed from multiple sources avoids requiring users to input individual bodies of text to our checker, instead automatically processing and detecting user votes for every news item. Instead of creating an entirely academic, infrequently used tool for news generation, we aim to offer a step toward integrating "fake news" detection within a typical news consumption routine, which could lead to wider adoption.

High-Quality Labelled Data Collection. A goal of our project was the ability to help collect high-quality labelled misinformation data. Through crowdsourcing credibility information from human feedback, we provide an automated data generation system. If several hundred human voters ranked an article as true, that information could in some cases be more useful than a single data labeler's assessment, especially when considering the decision-making fatigue which could result from labelling large amounts of data, as is often needed for machine learning model training. We sought to create a tool with results which could either be used directly to train machine learning models, or with "mostly accurate" labelled data which could accelerate the task of manual data labelling.

2.2. Evaluation of Tools and Frameworks

2.2.1 Frontend/Backend Language and Hosting

The CoVerifi frontend is a web app written in React.js, a JavaScript library created by Facebook. We used a sample React Twitter Feed web app, which was available on GitHub with a MIT license, allowing modification, distribution, and commercial use [75]. The frontend uses Firebase Hosting, which allows free hosting of React.js projects. Since we developed and host our own ML model, we used a Python-based architecture with Flask, a framework for developing web services that allows for the creation of API endpoints for communication between the backend and frontend. The backend is hosted by Heroku, which is free for our use case. Prior to deciding on Heroku, we attempted to host our service on AWS EC2 and AWS Elastic Beanstalk, but both had costs and complexities. A limitation, however, is that it sleeps after 30 minutes of inactivity [76], which means that the first API call within a 30-minute period will take 30 seconds, whereas subsequent calls will be quick.

2.2.2 News API Decision Process

Since the Google News Search API is deprecated, we chose from other existing options detailed in Table 1. Though we initially used newsapi.org due to their free tier and cheaper per-request cost, the Bing News Search API, has a free, 1000 requests/month option, and another option where every 1000 requests costs $4 (Table 1 provides a detailed breakdown of APIs and their advantages/disadvantages). While there exists seemingly cheaper options, such as ContextualWeb News API, we erred on the side of choosing a more widely-known tool, Bing News Search, since we plan on performing labelled data collection at a scale of between 1000 and 10,000 requests.

Service Name	Free requests	Paid tier pricing	Additional Info
NewsAPI.org [71]	500/day	$449 for 250,000, then $44.90 per 25,000 calls	Cannot make requests from browser in free tier
Bing News Search [72]	1,000/month	$4 per 1,000 transactions	Ability to set budgets. If 1,000/month is exceeded, no charge is incurred.
Currents API [73]	600/day	$150 for 300,000 requests, then $25 per 25,000 requests	Free tier says “No Access to Articles”, while paid tier says “Access to Articles”
ContextualWeb News API [74]	10,000/month	$0.5 for 1,000 requests after 10,000 exceeded	Not very well known. Potential for overage charges: “Depending on your plan's specification, you will either incur overage charges or be suspended.”

Table 1: Comparison of News APIs

2.2.3. Machine Learning Model

As a starting point, we implemented a text classification model trained on the outputs of OpenAI’s GPT-2 [32] model for generation of neural fake text, hosted by Hugging Face[2] as a part of their hosted inference API [77]. This free solution labels a piece of text as “fake” (meaning machine-generated) or “real” (meaning human-generated) with an associated confidence. Specific API endpoint information is provided in this paper’s GitHub repository at https://github.com/nlkolluri/CoVerifi. While we believe this model provides a valuable signal of whether text might be machine generated due to the intuition that text generated by different models may have similar characteristics, it is important to note that its accuracy is highest on fake text generated by GPT-2 versus other models. When discussing the GPT-2 output detector, the phrases “fake text” and “real text” only refer to whether the text was generated by a machine or by a human and do not in any way refer to whether the text’s content is considered true or false. As such, the GPT-2 output detector does not make a direct claim about the veracity of a piece of content. There are cases in which a language model produces a true sentence, and there are also cases in which a human writes a false statement. We chose to include the GPT-2 output detector in our platform because we believe that, while being machine generated is not the same as being false, knowing whether a piece of content is likely to be machine generated may help gauge credibility.

CoVerifi also features a machine-learning model which we developed and trained on a COVID-19 specific misinformation dataset, CoAID [78]. Moreover, our model and code usage samples are all available open source in order for other research teams to extend and develop

[2] We contacted Hugging Face about making calls to their API in our website and were told that this is permissible for academic use.

our work. We trained a Bidirectional LSTM on 1257 pieces of news content from CoAID and internally validated it on 419 pieces of news content from CoAID, as part of a 75% train and 25% test split. When using a weighted average across the two labels and rounding to the nearest hundredth, our F1-score was 0.93, with equal precision, recall, and accuracy. To assess generalizability to new types of COVID-19-specific misinformation, we created a dataset containing approximately 7,000 pieces of COVID-19-specific misinformation content. This dataset includes “fake news” labeled content (produced by Poynter) and verified news (for which we inherit news source credibility). Since the Poynter news contained several different labels, we assigned a label of 1 to all content with the label of “TRUE” and a label of 0 to all other content. As such, there may be a very small amount of mislabelled content present. We then tested our model trained on CoAID on this new dataset. When using a weighted average across the two labels and rounding to the nearest hundredth, our F1-score was 0.75, with equal precision, recall, and accuracy. While inheriting news source credibility may not be as reliable as manual labelling, this is less of an issue in an external validation set: it still shows that our model can perform better than random on new data sources. Furthermore, the F1- score for the “false” label, which came from Poynter, is 0.79. This is better than our F1-score of 0.70 for the “true” label.

Since our false-labelled news came from an established fact checking organization, Poynter, and our true-labelled news was obtained through the less-reliable method of inheriting news source credibility, it is reasonable to conclude that the false news F1-score is more indicative of the performance we would have with a perfectly-labelled dataset. This leads us to believe that rigorous manual labelling for each label in the external validation set could potentially increase accuracy.

2.2.4. Additional Tools Decision Process

Since News APIs typically only provide a brief subsection of the full article text along with a URL, we use the news-please [79] news crawler. To make the Twitter API easier to access, the Tweepy library was used [80]. We use Google Firebase’s Cloud Firestore, which allows for a document-based database.

3. System Design

As illustrated in Figure 1, CoVerifi can best be understood as 3 separate parts, (1) a frontend, public-facing web app written in React.js, and (2) a backend Python service which can be accessed through simple API calls from the front end with JSON input data and JSON output data, and (3) a database storing user vote information. Because they are separate parts, we hosted them using distinct services as detailed in Section 2.

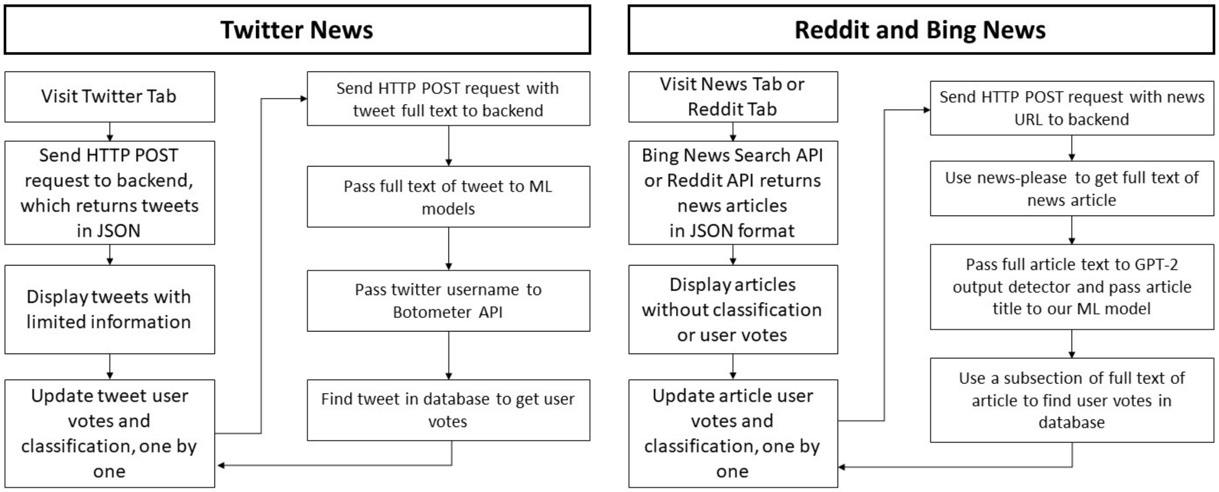

Twitter News
Visit Twitter Tab
Send HTTP POST request to backend, which returns tweets in JSON
Display tweets with limited information
Update tweet user votes and classification, one by one
Send HTTP POST request with tweet full text to backend
Pass full text of tweet to ML models
Pass twitter username to Botometer API
Find tweet in database to get user votes
Reddit and Bing News
Visit News Tab or Reddit Tab
Bing News Search API or Reddit API returns news articles in JSON format
Display articles without classification or user votes
Update article user votes and classification, one by one
Send HTTP POST request with news URL to backend
Use news-please to get full text of news article
Pass full article text to GPT-2 output detector and pass article title to our ML model
Use a subsection of full text of article to find user votes in database

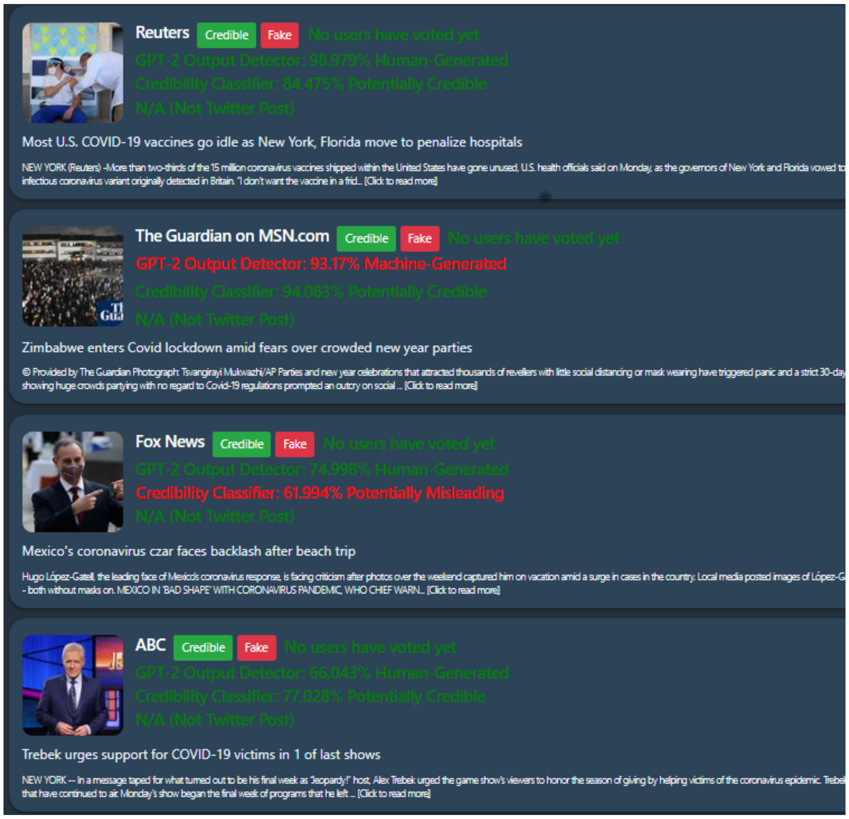
Figure 1: System Interactions of CoVerifi

3.1. CoVerifi Frontend

CoVerifi has 5 different news options: COVID-19 news from the Bing News Search API, Breaking News from the Bing News Search API, Reddit's news subreddit, news from Twitter, and the option to "search" for a specific query in the Bing News Search API by using the search bar. For each of these options, a "feed" is created with several entries, each corresponding to a news article, Reddit post, or tweet and information about the piece of news content, the machine learning model assessments of the content's credibility, the credibility of the content as assigned by the user voters, a button to rate the news as "credible", and a button to rate the news as "fake" is displayed (see figures 2-3).

3.1.1. Bing News Search API and Reddit News

For the Bing News Search API and the Reddit news API, the respective API is called with the query information. This leaves limited information about the article, since the Bing News Search API and Reddit API only load very limited content from the article/post itself. Thus, for each entry, the URL of each piece of content is sent to the backend API in JSON format. The backend returns the full-text of the article, the GPT-2 output detector model's classification of how likely the piece of content is real or fake, and our machine learning model's classification of the title of the news content in JSON format. The pieces of content are updated one-by-one as the backend finishes processing them. Given that the Heroku backend "sleeps" (causes a 30 second response time after 30 minutes inactivity), constructing the frontend in this way allowed us to mimic the responsive behavior of our code at production. The Bing News API requests can eventually originate from the backend.



Reuters Credible Fake No users have voted yet
GPT-2 Output Detector: 98.979% Human-Generated
Credibility Classifier: 84.475% Potentially Credible
N/A (Not Twitter Post)

Most U.S. COVID-19 vaccines go idle as New York, Florida move to penalize hospitals

NEW YORK (Reuters) -More than two-thirds of the 15 million coronavirus vaccines shipped within the United States have gone unused, U.S. health officials said on Monday, as the governors of New York and Florida vowed to
infectious coronavirus variant originally detected in Britain. 'I don't want the vaccine in a frid... [Click to read more]

The Guardian on MSN.com Credible Fake No users have voted yet
GPT-2 Output Detector: 93.17% Machine-Generated
Credibility Classifier: 94.083% Potentially Credible
N/A (Not Twitter Post)

Zimbabwe enters Covid lockdown amid fears over crowded new year parties

© Provided by The Guardian Photograph: Tsvangirayi Mukwazhi/AP Parties and new year celebrations that attracted thousands of revellers with little social distancing or mask wearing have triggered panic and a strict 30-day
showing huge crowds partying with no regard to Covid-19 regulations prompted an outcry on social ... [Click to read more]

Fox News Credible Fake No users have voted yet
GPT-2 Output Detector: 74.998% Human-Generated
Credibility Classifier: 61.994% Potentially Misleading
N/A (Not Twitter Post)

Mexico's coronavirus czar faces backlash after beach trip

Hugo López-Gatell, the leading face of Mexico's coronavirus response, is facing criticism after photos over the weekend captured him on vacation amid a surge in cases in the country. Local media posted images of López-G
- both without masks on. MEXICO IN 'BAD SHAPE' WITH CORONAVIRUS PANDEMIC, WHO CHIEF WARN... [Click to read more]

ABC Credible Fake No users have voted yet
GPT-2 Output Detector: 66.043% Human-Generated
Credibility Classifier: 77.028% Potentially Credible
N/A (Not Twitter Post)

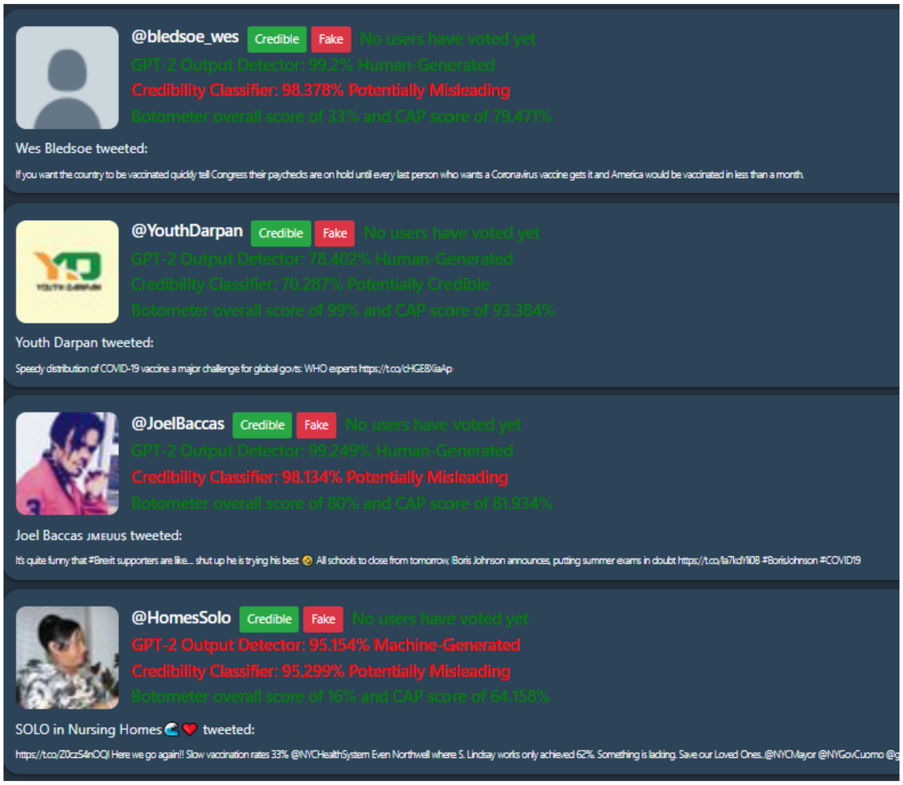
Trebek urges support for COVID-19 victims in 1 of last shows

NEW YORK -- In a message taped for what turned out to be his final week as "Jeopardy!" host, Alex Trebek urged the game show's viewers to honor the season of giving by helping victims of the coronavirus epidemic. Trebek
that have continued to air. Monday's show began the final week of programs that he left ... [Click to read more]

Figure 2: COVID-19 News from Bing News Search API

3.1.2. Twitter News

For Twitter, a different format is followed wherein the frontend makes a call to the backend with the Twitter query (i.e, 'COVID-19' or 'Coronavirus'). The backend returns a JSON object with the information to be displayed about several tweets. Articles were not classified in the prior step. Then, each tweet is individually passed to the backend to perform classification by the machine learning models.



@bledsoe_wes Credible Fake No users have voted yet
GPT-2 Output Detector: 99.2% Human-Generated
Credibility Classifier: 98.378% Potentially Misleading
Botometer overall score of 33% and CAP score of 79.471%

Wes Bledsoe tweeted:

If you want the country to be vaccinated quickly tell Congress their paychecks are on hold until every last person who wants a Coronavirus vaccine gets it and America would be vaccinated in less than a month.

@YouthDarpan Credible Fake No users have voted yet
GPT-2 Output Detector: 78.402% Human-Generated
Credibility Classifier: 70.287% Potentially Credible
Botometer overall score of 99% and CAP score of 93.384%

Youth Darpan tweeted:

Speedy distribution of COVID-19 vaccine a major challenge for global govts: WHO experts https://t.co/cHGE8XiaAp

@JoelBaccas Credible Fake No users have voted yet
GPT-2 Output Detector: 99.249% Human-Generated
Credibility Classifier: 98.134% Potentially Misleading
Botometer overall score of 80% and CAP score of 81.934%

Joel Baccas ᴊᴍᴇᴜᴜs tweeted:

It's quite funny that #Brexit supporters are like... shut up he is trying his best 🤣 All schools to close from tomorrow, Boris Johnson announces, putting summer exams in doubt https://t.co/la7kdYlI0B #BorisJohnson #COVID19

@HomesSolo Credible Fake No users have voted yet
GPT-2 Output Detector: 95.154% Machine-Generated
Credibility Classifier: 95.299% Potentially Misleading
Botometer overall score of 16% and CAP score of 64.158%

SOLO in Nursing Homes🌊❤️ tweeted:

https://t.co/Z0cz54nOQI Here we go again!! Slow vaccination rates 33% @NYCHealthSystem Even Northwell where S. Lindsay works only achieved 62%. Something is lacking. Save our Loved Ones..@NYCMayor @NYGovCuomo @g

Figure 3: Tweets from Twitter API

3.1.3. User Credibility Ratings

To display the user credibility ratings on the initial load of the page, the database is checked for each entry, using the expanded text of the article or tweet as the database key. Once a user votes on a particular entry, their vote is recorded and the updated state of the database is retrieved. This means that if a second user voted between the time that the page loaded and when the first user voted, the second user's vote is still displayed. Displaying the most recent state of the database without needing to refresh the page thus functions as an incentive for the user to vote, which in turn allows us to collect more labelled data with respect to the credibility of news content.

To combat concerns about abuse of crowd-sourced credibility ratings, we have implemented IP-based abuse prevention. There can only be one vote per piece of content per IP address. The IP-based vote limiting is accomplished by retrieving the IP address on the client side using ipapi.co [81] and only allowing that IP address to cast one vote per piece of content. It is important to note that the current implementation of our crowd-based labeling in CoVerifi would benefit from improvements in security (see Section 5 for a detailed discussion of limitations and areas for improvement).

3.2. CoVerifi Backend

Our design consists of three endpoints: one for getting tweets, one for classifying tweets, and one for getting article expanded text and processing the article, all of which can be accessed through a HTTP POST request.

Collecting and Processing Tweets. Once the information queried from the Twitter API via the Tweepy package is received, tweets are filtered to ensure that they are not retweets. A custom JSON object is created which includes selected information from every tweet returned by Tweepy. Then, the JSON object is returned as an HTTP response and used by the frontend. The endpoint for processing tweets is separate from the collecting tweets endpoint so that individual tweets can be processed one by one, rather than delaying the display of tweets until all tweets have been processed. The expanded text of the tweet is sent to the endpoint with an HTTP POST request in the form of a JSON object. Then, a classifier trained on outputs from OpenAI's GPT-2 for detection of neural "fake news" is used to process the tweet, which in practice means submitting a HTTP request to an API provided by Hugging Face [77]. At this step, we also collect the Botometer score for the tweet's poster. These metrics for determining if the tweet is credible, wrapped in a JSON object, are returned as an HTTP response.

Get Article Text and Process. The endpoint for getting article expanded text and processing article expanded text is accessed through an HTTP POST request containing a JSON object with the URL of the piece of news content. The URL is passed to the news-please library, which retrieves the entire text of the news article being processed. A 300-character subsection of the expanded text of the article is then passed into the GPT-2 output detector model API endpoint for neural "fake news" detection. We also pass the title of the article into our machine learning model, which outputs a credibility score.

The expanded text of the article, the credibility rating outputted by the GPT-2 output detector model, and the output of our machine learning model are packaged together in a JSON object and returned as an HTTP Response. The frontend uses the updated expanded text to retrieve information from the database, as well as to display more information from the article where relevant.

Database. CoVerifi's Cloud Firestore database consists of a collection of documents, where the key of each document is a subsection of the expanded text of a piece of news content, which allows each document to be uniquely identified in the database. An alternative keying strategy could be using the URL of the piece of news content, but the current strategy allows our web app to be expanded to display types of news content which can be collected via API call but may not have an associated link, or content with a non-constant URL.

4. Discussion and Implications for Future Work

CoVerifi enables many future research directions as its code is provided as open source. The source code is available at https://github.com/nlkolluri/CoVerifi and the website is publicly available at https://coverifi.web.app/. This section reflects on the utility of CoVerifi and what lines of research the platform opens up for future work.

4.1. Misinformation Hypothesis Testing

A survey of 2501 respondents from Singapore found that most social media users simply ignore "fake news" posts they come across on social media, only offering corrections when the issue is strongly relevant to them and to people with whom they share a strong personal relationship [82]. An active call for intervention is potentially useful because it would allow researchers to frame the COVID-19 infodemic as something strongly relevant to media consumers, which in turn could allow organic checks on misinformation. CoVerifi fulfills this function by actively asking users to provide their feedback on the credibility of news content which is displayed on CoVerifi. Second, by providing users an abundance of potentially false news information which needs classification, CoVerifi reinforces the notion that “fake news” is an issue which requires intervention on behalf of anyone who encounters it. This would enable researchers to compare the truth discernment on CoVerifi, where an active call for intervention is requested, with the truth discernment of users deciding what to share on social media. Additionally, research could be done comparing a group who first used CoVerifi and then used social media with a control group who only used social media. It is possible that prior usage of CoVerifi could have a priming effect and result in media consumers being more likely to intervene against misinformation when they encounter it.

Given that when participants were primed to think about accuracy at the beginning of a study, their level of truth discernment in whether they intended to share the COVID-19-related articles was more than doubled [61], a future direction for use with our platform is quantifying the increase in effectiveness at identifying “fake news” which users experience when using our platform. Perhaps asking users to vote on individual pieces of news will implicitly prime users to consider accuracy in their responses. It could also be examined whether using CoVerifi, combined with explicit truth nudging, yields any benefit over using just CoVerifi. Evaluating the efficacy of either subtly or overtly nudging users to think about the accuracy of content is another potential use case of CoVerifi.

Because users on mainstream media platforms are often at most a few clicks away from malicious, hateful content [9], one line of research could be to evaluate the effectiveness of navigation warnings at preventing users from ending up on a page with malicious, hateful misinformation. Due to the demonstrated effectiveness at truth-nudging [61], it is possible that giving a warning prior to a user leaving CoVerifi for another, potentially unreliable site could increase user awareness of unreliable content.

With the presence of malicious COVID-19 narratives being potentially linked to offline anti-Asian violence [5] [9] [10], hateful misinformation is a real issue. Perhaps an approach which combines traditional credibility detection strategies with sentiment detection strategies to identify hateful or aggressive sentiment could be useful in helping prevent the spread of hateful misinformation. In the context of our web app, a sentiment rating could be displayed below the computed credibility rating. This would allow research on whether angry or aggressive sentiment is correlated with malicious misinformation.

4.2. Optimization and Scaling

Through moving to use a more scalable hosting set-up, a different machine learning model, and different news API payment tier (or implementing a query caching system), we aim to increase the scalability of CoVerifi. This would allow seamless integration of fact verification with the average media diet and news consumption routines, allowing us to help reduce the spread of COVID-related misinformation.

In the context of our project, we could continually update the machine learning model used on our CoVerifi as we collect more data. Through combining machine learning with human feedback, we can minimize the errors present in either approach. Due to the fact that “fake news” often includes a confusing mix of “true” news and misinformation, humans can often be inaccurate at identifying “fake news” on their own, achieving only 50-63% success at identifying “fake news” [83]. Another study found that while humans were better at identifying “fake news” content in the Celebrity category than the automated system, their system outperformed humans while detecting “fake news” in more serious and diverse news sources [55]. Due to human error in “fake news” identification, it is possible that our display of a machine learning model’s classification prior to a user voting may minimize the rate of incorrect human classification. The result may be that our crowdsourced data is higher quality than the initial labelled dataset, and this hypothesis could be tested by an iterative approach of continually retraining as more data is introduced. Moreover, due to the established presence of hateful and malicious information on a variety of platforms, there remains a need for ways to check the validity of content on new platforms. CoVerifi could be used to rapidly develop new content consumption streams by simply swapping the API from Reddit/Twitter to an API which collects content from a new information sharing platform.

4.3. Data Collection and Analysis

Given that models trained against a crowdsourced dataset (CREDBANK) outperformed models trained against the journalists’ assessed dataset (PHEME) when tested on a set of credibility-labelled tweets [56], crowdsourcing data may have considerable merit. Our web app can thus function as an easy-to-use interface to allow crowdsourcing of factuality data, which could in turn help create high-quality labelled datasets. While we believe crowdsourced credibility assessments are a useful feature of the CoVerifi platform, this approach is not new. A decade ago, Ratkiewicz et al. [84] used crowdsourced judgements to label memes as “Truthy” or “Legitimate”.

Since CoVerifi leverages multiple platforms (Bing News Search, Reddit, Twitter) to provide a diverse set of news providers mimicking a typical media diet, our platform could be used to perform comparative analysis on how well machine learning models respond to the various platforms. Additionally, our tool could allow comparative research on how users perceive the accuracy of platforms.

We evaluated whether training on only COVID-19- specific “fake news”, on only general “fake news”, or on a combination of COVID-19-specific “fake news” and general “fake news” would yield the best results when tested on a new COVID-19-specific dataset. We used a COVID-19-specific “fake news” dataset from CoAID and a general “fake news” dataset from FakeNewsNet to evaluate generalizability to our dataset discussed in Section 2.2.3. For 7 different combinations of data sources including COVID-19 “fake news”, general “fake news”, or both, we trained implementations of a Support Vector Machine (SVM) classifier, a Logistic Regression classifier, and a Bernoulli Naive Bayes (Bernoulli NB) model. We found that the inclusion of CoAID maintained or improved performance compared to not including it, confirming that including COVID-19-specific misinformation content improves model performance.

5. Limitations

Our project does have limitations as many design decisions were significantly influenced by prioritizing speed and cost. Due to CoVerifi’s lack of set-up costs, it is not yet suited for large-scale usage. This limits the scale of research questions that can be answered using our tool in its present state. Specifically, Heroku’s hosting will only allow us to have our server “awake” for 550 hours/month, sleeping after 30 minutes of activity. Additionally, we are currently using the Bing News API in the 1000 requests/month setting, but in the future, we will either upgrade to a setting allowing additional requests at $4 per 1000 requests or perform caching of search results to decrease the number of API calls. Moreover, a limitation of our model choice is that it only detects robot-generated “fake news”, though this opens up future directions for model development that we discuss in Section 4. Lastly, CoVerifi's current news inputs reflect predominantly Western preferences; however, other, more international API endpoints can be added as needed by others.

While CoVerifi cannot prevent certain platforms from eventually restricting access to their content, it appears that the presence of multiple avenues for obtaining content (i.e., news sources pulled from Bing News Search, Reddit news, and Twitter) will prevent CoVerifi from becoming obsolete. We also believe that CoVerifi, through providing multiple credibility metrics in a single unified location, minimizes the amount of time required by the user. However, there is a necessity of user interpretation, since prior truth nudging research [61] indicates that critically thinking about accuracy is beneficial for truth discernment.

A significant limitation of CoVerifi is that our crowd-based labelling is not currently at a fully secure stage. Specifically, CoVerifi could be vulnerable to a coordinated attack aimed at deceptively enhancing the credibility of a specific news feed. While CoVerifi protects against a single user repeatedly voting on a piece of content through our website from the same IP address, it does not protect against more sophisticated approaches involving programmatically accessing our database, coordinated attacks performed by several users, or a user voting from several different IP addresses. One important direction for increasing the security of our crowd-based labelling is handling all database access on the backend rather than the frontend to decrease the vulnerability of our data to programmatic manipulation. Additionally, IP-based activity monitoring and analysis should be used to detect suspicious behavior. For example, if sets of IP addresses always tend to promote the same content or if sets of IP addresses always disagree with our machine learning models, these IPs could be flagged for manual review. Such approaches have the potential to mitigate the risk associated with both coordinated attacks performed by several users as well as attacks performed by single users across several IP addresses. The IP-based activity monitoring functionality might also be used to cross-check all votes. If a vote has been recorded in the database without a valid IP address from which it originated, this could reveal vulnerabilities in the crowd-based labelling mechanism and indicate that the integrity of the crowd-based data has been compromised. Another option for improving the security of our system is to incorporate user authentication and require users to be logged in to vote. This could be combined with IP checking and the blocking of suspicious user behavior. Given that our crowd-based labelling is not completely secure, it is important that future researchers are mindful of limitations and consider how to incorporate a level of security appropriate for their use case.

Furthermore, there is a concern highlighted in the literature that a labelled news-feed could backfire by making users too dependent on the credibility labels, unable to discern truth for themselves. For example, in the case of detecting images related to "fake news", misinformation, disinformation, credibility labels were found to be "supplanting reasoned deliberation with mechanistic verification" [85]. CoVerifi addresses this concern by providing (1) a disclaimer section and (2) a multi-faceted automated credibility detection approach. The disclaimer section of CoVerifi provides accuracy metrics for our machine learning model in order to establish that the possibility of error is very real and we are transparent in this disclosure. In our disclaimer, we have included text to indicate that readers should be critical of the news content they read. This is drawn from other empirical work which shows that truth-nudging increases the accuracy of a user's own critical judgements [61].

We also utilize a multi-faceted automated credibility detection approach. For news content, the neural "fake news" detector can give a signal that the text may have been machine-generated and the machine learning model we trained can provide a signal that the news article title is similar to that of "fake news" articles. For Twitter, the neural detector can produce a signal that the text itself may have been machine-generated, while the Botometer API score can produce a signal that the account posting the tweets may be a bot. Through providing multiple credibility scores which often disagree with each other, we highlight the active role of the user to decide the credibility of the piece of news content they are viewing. CoVerifi critically provides additional information and adds a data-driven layer to a user's ability to discern credibility, rather than asserting a definitive credibility score. A user can employ this information to help decide whether a seemingly-robotic piece of content may have actually been generated by a robot, whether a hyperbolic title bears similarity to "fake news", or whether other viewers of the site think the content is false.

The literature also raises concerns that credibility labels could simply be ineffective. While Gao et al. [86] "did not find that credibility labels had any effect on people's perception of fake news", they found that "credibility labels mitigate selective exposure bias, especially for users with liberal stances", which suggests that "credibility labels could marginally decrease people's level of agreement on news articles on their own side, which may lead to a more moderate opinion space". However, Gao et al.'s [86] study only contained content on two topics (gun control and U.S. President Donald Trump) from 14 articles (8 of which were labelled for credibility) and acknowledges the benefits of more long-term studies on the impact of labelling. The limitations of their study, the lack of significant negative effects found from credibility labelling, and the impact of credibility labelling on mitigating selective exposure bias indicates that further work which evaluates credibility labeling remains worthwhile.

Other work indicates that repeatedly labelling a claim as false can given the illusion that the claim is true by increasing familiarity with the content [87]. This is supported by studies that have demonstrated that exposing people to claims increases the perceived truth of the claim when it is seen again later [87] [88] [89]. This is the case even for statements that are explicitly identified as false on initial presentation [87] [90] [91]. As Polage, argues, news source credibility has been found to be directly affected by repeated exposure [92]. Specifically, "people will believe information to be true if it is repeated, if it does not contradict previously stored knowledge, and if the source has not been discredited" [92]. CoVerifi mitigates this issue due to the many unique pieces of content displayed on the platform at any given time. Since we derive content from the APIs of Bing News, Reddit, and Twitter dynamically, the volume and variety of content means that the risk of familiarity with content is very low.

6. Conclusion

The explosion of misinformation, disinformation and hate news associated with the COVID-19 infodemic has left fact checkers overburdened. Given the massive quantity of "fake news", automated approaches are imminently important as a way of minimizing the damage of the infodemic. We introduce CoVerifi, a solution which combines the power of truth-nudging, human feedback, and machine learning in a highly platform- and information-agnostic manner. CoVerifi provides a multi-channel credibility check for news, which means that "fake news" that has failed to be detected by the automated approaches can be detected by user feedback and vice versa. The presence of multiple platforms means that CoVerifi can reflect diverse media diets. Moreover, this enables CoVerifi to be used to analyze new types of content on more regionally specific platforms. Our open sourced code enables others to host CoVerifi, including deploying it with paid services for greater scalability. CoVerifi's code could then be connected to a different classification model or a different news API to analyze different types of data.

Furthermore, the querying feature present in the current version allows virtually any news article to be fact checked, since Bing News Search contains articles from an expansive range of news sources. It is our intention that CoVerifi provides a starting point for new research directions, allowing researchers to rapidly create accessible services to address a wide range of misinformation concerns and research questions across a broad spectrum of platforms and disciplines.

7. Declaration of Competing Interests

The authors declare that there are no competing interests.

8. Acknowledgements

The work presented in this paper was partially funded by the Moody College of Communication, University of Texas at Austin.

References

[1] J. Hua, R. Shaw, Corona virus (covid-19) “infodemic” and emerging issues through a data lens: The case of china, International Journal of Environmental Research and Public Health 17 (7) (2020) 2309. doi:10.3390/ijerph17072309. URL http://dx.doi.org/10.3390/ijerph17072309

[2] World Health Organization, Coronavirus disease (COVID-19) pandemic (2020 (accessed November 23, 2020)). URL https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[3] C. M. Pulido, B. Villarejo-Carballido, G. Redondo-Sama, A. Gómez, Covid-19 infodemic: More retweets for science-based information on coronavirus than for false information, International Sociology 35 (4) (2020) 377–392. arXiv: https://doi.org/10.1177/0268580920914755, doi:10.1177/0268580920914755. URL https://doi.org/10.1177/0268580920914755

[4] J. Brennen, F. Simon, P. Howard, R. Nielsen, Types, sources, and claims of covid-19 misinformation (04 2020).

[5] R. F. Sear, N. Velásquez, R. Leahy, N. J. Restrepo, S. E. Oud, N. Gabriel, Y. Lupu, N. F. Johnson, Quantifying covid-19 content in the online health opinion war using machine learning, IEEE Access 8 (2020) 91886–91893.

[6] S. Almasy, H. Yan, M. Holcombe, Coronavirus Pandemic Hitting Some African-American Communities Extremely Hard (2020 (accessed July 18, 2020)). URL https://www.cnn.com/2020/04/06/health/us-coronavirus-updates-monday/index.html

[7] A. Maqbool, Coronavirus: Why has the virus hit African Americans so hard? (2020 (accessed July 18, 2020)). URL https://www.bbc.com/news/world-us-canada-52245690

[8] A. R. Lab, APM Research Lab (2020 (Accessed November 23, 2020)). URL https://www.apmresearchlab.org/

[9] N. Velásquez, R. Leahy, N. J. Restrepo, Y. Lupu, R. Sear, N. Gabriel, O. Jha, B. Goldberg, N. F. Johnson, Hate multiverse spreads malicious covid-19 content online beyond individual platform control (2020). arXiv:2004.00673.

[10] H. Yan, N. Chen, D. Naresh, What's spreading faster than coronavirus in the US? Racist assaults and ignorant attacks against Asians (2020 (accessed July 18, 2020)). URL https://www.cnn.com/2020/02/20/us/coronavirus-racist-attacks-against-asian-americans/index.html

[11] M. Sayyadiharikandeh, O. Varol, K.-C. Yang, A. Flammini, F. Menczer, Detection of novel social bots by ensembles of specialized classifiers (2020). arXiv:2006.06867.

[12] IUNetSci, Botometer Python API (2020 (Accessed November 15, 2020)). URL https://github.com/IUNetSci/botometer-python

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

[14] E. C. Tandoc, Jr., Z. W. Lim, R. Ling, Defining "fake news", Digital Journalism 6 (2) (2018) 137–153. arXiv:https: //doi.org/10.1080/21670811.2017.1360143, doi:10.1080/21670811.2017.1360143. URL https://doi.org/10.1080/21670811.2017.1360143

[15] A. Gelfert, Fake news: A definition, Informal Logic 38 (1) (2018) 84–117. doi: https://doi.org/10.22329/il.v38i1.5068.

[16] K. Dalkir, R. Katz, Navigating Fake News, Alternative Facts, and Misinformation in a Post-Truth World, Advances in Media, Entertainment, and the Arts, IGI Global, 2020. URL https://books.google.com/books?id=RLzTDwAAQBAJ

[17] C. Wardle, Fake news. It's complicated. (2017 (accessed July 18, 2020)). URL https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79

[18] G. L. Ciampaglia, Fighting fake news: a role for computational social science in the fight against digital misinformation, Journal of Computational Social Science (2018) 147–153. doi: 10.1007/s42001-017-0005-6.

[19] G. L. Ciampaglia, A. Mantzarlis, G. Maus, F. Menczer, Research challenges of digital misinformation: Toward a trustworthy web, AI Magazine 39 (1) (2018) 65–74. doi:10.1609/aimag.v39i1.2783. URL https://www.aaai.org/ojs/index.php/aimagazine/article/view/2783

[20] D. Nikolov, D. Oliveira, A. Flammini, F. Menczer, Measuring online social bubbles, PeerJ Computer Science 1 (02 2015). doi:10.7717/peerj-cs.38.

[21] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, A. Scala, The covid-19 social media infodemic (2020). arXiv:2003.05004.

[22] E. Ferrara, Disinformation and social bot operations in the run up to the 2017 french presidential election, First Monday 22 (8) (Jul 2017). doi:10.5210/fm.v22i8.8005. URL http://dx.doi.org/10.5210/fm.v22i8.8005

[23] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, F. Menczer, Botornot: A system to evaluate social bots, in: Proceedings of the 25th International Conference Companion on World Wide Web, WWW ’16 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2016, p. 273–274. doi:10.1145/2872518.2889302. URL https://doi.org/10.1145/2872518.2889302

[24] J. P. Robinson, How Americans Use Time: A Social-psychological Analysis of Everyday Behavior, Praeger scientific, Praeger, 1977. URL https://books.google.com/books?id=UL3ZAAAAMAAJ

[25] J. Kulshrestha, M. Zafar, L. Noboa, K. Gummadi, S. Ghosh, Characterizing information diets of social media users (2015). URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10595/10505

[26] J. Ohme, When digital natives enter the electorate: Political social media use among first-time voters and its effects on campaign participation, Journal of Information Technology & Politics 16 (2) (2019) 119–136. arXiv: https://doi.org/10.1080/19331681.2019.1613279, doi:10.1080/19331681.2019.1613279. URL https://doi.org/10.1080/19331681.2019.1613279

[27] DeepAI, Machine Learning (n.d. (Accessed 7/18/2020)). URL https://deepai.org/machinelearning-glossary-and-terms/machinelearning

[28] Z. Zhou, H. Guan, M. Bhat, J. Hsu, Fake news detection via nlp is vulnerable to adversarial attacks, Proceedings of the 11th International Conference on Agents and Artificial Intelligence (2019). doi:10.5220/0007566307940800. URL http://dx.doi.org/10.5220/0007566307940800

[29] Fake News Challenge, FREQUENTLY ASKED QUESTIONS (2017 (Accessed July 18, 2020)). URL http://www.fakenewschallenge.org/

[30] I. Turc, M.-W. Chang, K. Lee, K. Toutanova, Well-read students learn better: On the importance of pre-training compact models (2019). arXiv:1908.08962.

[31] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi, Defending against neural fake news (2019). arXiv: 1905.12616.

[32] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019.

[33] OpenAI, GPT-2 Output Detector (2019 (accessed July 18, 2020)). URL https://github.com/openai/gpt-2-output-dataset/tree/master/detector

[34] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fact extraction and VERification (FEVER) shared task, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels,

Belgium, 2018, pp. 1–9. doi:10.18653/v1/W18-5501. URL https://www.aclweb.org/anthology/W18-5501

[35] ClaimBuster, Fact Checker (n.d. (accessed July 18, 2020)). URL https://idir.uta.edu/claimbuster/factchecker/

[36] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, V. Sable, C. Li, M. Tremayne, Claimbuster: The first-ever end-to-end fact-checking system, Proc. VLDB Endow. 10 (12) (2017) 1945–1948. doi:10.14778/3137765.3137815. URL https://doi.org/10.14778/3137765.3137815

[37] Google, Fact Check Explorer (n.d. (accessed July 18, 2020)). URL https://toolbox.google.com/factcheck/explorer

[38] S. Gehrmann, H. Strobelt, A. M. Rush, Gltr: Statistical detection and visualization of generated text (2019). arXiv:1906.04043.

[39] H. Strobelt, S. Gehrmann, Catching a Unicorn with GLTR: A tool to detect automatically generated text (2019 (Accessed July 18, 2020)). URL http://gltr.io/

[40] S. O. for Disinformation, S. M. Analysis, Homepage, SOMA Disinfobservatory (2020 (Accessed November 15, 2020)). URL https://www.disinfobservatory.org/

[41] S. Guarino, N. Trino, A. Celestini, A. Chessa, G. Riotta, Characterizing networks of propaganda on twitter: a case study, Applied Network Science 5 (1) (Sep 2020). doi:10.1007/s41109-020-00286-y. URL http://dx.doi.org/10.1007/s41109-020-00286-y

[42] S. Guarino, N. Trino, A. Chessa, G. Riotta, Beyond Fact-Checking: Network Analysis Tools for Monitoring Disinformation in Social Media, 2019, pp. 436–447. doi:10.1007/978-3-030-36687-2_36.

[43] L. Toumanidis, R. Heartfield, P. Kasnesis, G. Loukas, C. Patrikakis, A Prototype Framework for Assessing Information Provenance in Decentralised Social Media: The EUNOMIA Concept, 2020, pp. 196–208. doi:10.1007/978-3-030-37545-4_13.

[44] EUNOMIA, The Project, Eunomia (2020 (Accessed November 15, 2020)). URL https://www.eunomia.social/project

[45] SocialTruth, The SocialTruth Project (2020 (Accessed November 15, 2020)). URL http://www.socialtruth.eu/

[46] M. Choras, M. Pawlicki, R. Kozik, K. Demestichas, P. Kosmides, M. Gupta, Socialtruth project approach to online disinformation (fake news) detection and mitigation, in: Proceedings of the 14th International Conference on Availability, Reliability and Security, ARES ’19, Association for Computing Machinery, New York, NY, USA, 2019. doi:10.1145/3339252.3341497. URL https://doi.org/10.1145/3339252.3341497

[47] WeVerify, About us - WeVerify (2020 (Accessed November 15, 2020)). URL https://weverify.eu/about/

[48] Z. Marinova, J. Spangenberg, D. Teyssou, S. Papadopoulos, N. Sarris, A. Alaphilippe, K. Bontcheva, Weverify: Wider and enhanced verification for you project overview and tools, in: 2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW), 2020, pp. 1–4. doi:10.1109/ICMEW46912.2020.9106056.

[49] WeVerify, Verification plugin - WeVerify (2020 (Accessed November 15, 2020)). URL https://weverify.eu/verificationplugin/

[50] Provenance, About Provenance (2020 (Accessed November 15, 2020)). URL https://www.provenanceh2020.eu/ about/about-provenance

[51] E. Culloty, J. Suiter, Beyond fact-checking: Countering the spread of political disinformation, The European Consortium for Political Research, Wroc law, Poland, 2019. URL https://ecpr.eu/Events/Event/PaperDetails/46889

[52] W. Y. Wang, "liar, liar pants on fire": A new benchmark dataset for fake news detection, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 422–426. doi:10.18653/v1/P17-2067. URL https://www.aclweb.org/anthology/P17-2067

[53] H. Ahmed, I. Traore, S. Saad, Detecting opinion spams and fake news using text classification, Security and Privacy 1 (1) (2018) e9. arXiv: https://onlinelibrary.wiley.com/doi/pdf/10.1002/spy2.9, doi:10.1002/spy2.9. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/spy2.9

[54] I. Traore, S. Saad, Detection of online fake news using n-gram analysis and machine learning techniques, in: Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, 2017, pp. 127–138. doi:10.1007/978-3-319-69155-8_9.

[55] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3391–3401. URL https://www.aclweb.org/anthology/C18-1287

[56] C. Buntain, J. Golbeck, Automatically identifying fake news in popular twitter threads, in: 2017 IEEE International Conference on Smart Cloud (SmartCloud), 2017, pp. 208–215.

[57] F. T. Asr, M. Taboada, Big data and quality data for fake news and misinformation detection, Big Data & Society 6 (1) (2019) 2053951719843310. arXiv: https://doi.org/10.1177/2053951719843310, doi:10.1177/2053951719843310. URL https://doi.org/10.1177/2053951719843310

[58] J. Bullock, A. Luccioni, K. H. Pham, C. S. N. Lam, M. Luengo-Oroz, Mapping the landscape of artificial intelligence applications against covid-19 (2020). arXiv:2003.11336.

[59] R. Kouzy, J. A. Jaoude, A. Kraitem, M. B. E. Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, K. Baddour, Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter, Cureus 12 (03 2020). doi:10.7759/cureus.7255.

[60] J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng, X. Liu, T. Zhu, Twitter discussions and emotions about covid-19 pandemic: a machine learning approach (2020). arXiv:2005.12830.

[61] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, D. G. Rand, Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention, Psychological Science 31 (7) (2020) 770–780, pMID: 32603243. arXiv: https://doi.org/10.1177/0956797620939054, doi:10.1177/0956797620939054. URL https://doi.org/10.1177/0956797620939054

[62] R. Pandey, V. Gautam, C. Jain, P. Syal, H. Sharma, K. Bhagat, R. Pal, L. S. Dhingra, Arushi, L. Patel, M. Agarwal, S. Agrawal, M. Arora, B. Rana, P. Kumaraguru, T. Sethi, A machine learning application for raising wash awareness in the times of covid-19 pandemic (2020). arXiv:2003.07074.

[63] A. Bruns, S. Harrington, E. Hurcombe, ‘corona? 5g? or both?’: the dynamics of covid-19/5g conspiracy theories on facebook, Media International Australia 177 (1) (2020) 12–29. arXiv: https://doi.org/10.1177/1329878X20946113, doi:10.1177/1329878X20946113. URL https://doi.org/10.1177/1329878X20946113

[64] Y. Mejova, K. Kalimeri, Covid-19 on facebook ads: Competing agendas around a public health crisis, in: Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS ’20, Association for Computing Machinery, New York, NY, USA, 2020, p. 22–31. doi:10.1145/3378393.3402241. URL https://doi.org/10.1145/3378393.3402241

[65] A. Celestini, M. D. Giovanni, S. Guarino, F. Pierri, Information disorders on italian facebook during covid-19 infodemic (2020). arXiv: 2007.11302.

[66] S. Boberg, T. Quandt, T. Schatto-Eckrodt, L. Frischlich, Pandemic populism: Facebook pages of alternative news media and the corona crisis – a computational content analysis (2020). arXiv:2004.02566.

[67] S. A. Memon, K. M. Carley, Characterizing covid-19 misinformation communities using a novel twitter dataset (2020). arXiv:2008.00791.

[68] E. Ferrara, What types of covid-19 conspiracies are populated by twitter bots?, First Monday (May 2020). doi:10.5210/fm.v25i6.10633. URL http://dx.doi.org/10.5210/fm.v25i6.10633

[69] L. Singh, L. Bode, C. Budak, K. Kawintiranon, C. Padden, E. Vraga, Understanding high- and low-quality url sharing on covid-19 twitter streams, Journal of Computational Social Science 3 (2020) 343–366. doi:10.1007/s42001-020-00093-6.

[70] B. Huang, K. M. Carley, Disinformation and misinformation on twitter during the novel coronavirus outbreak (2020). arXiv:2006.04278.

[71] NewsAPI, Pricing (n.d. (accessed July 18, 2020)). URL https://newsapi.org/pricing

[72] Microsoft, Cognitive Services Pricing – Bing Search API (n.d. (Accessed July 18, 2020)). URL https://azure.microsoft.com/en-us/pricing/details/cognitive-services/search-api/

[73] CurrentsAPI, PRICING OVERVIEW (n.d. (accessed July 18, 2020)). URL https://currentsapi.services/en/product/price

[74] contextualwebsearch, Contextual Web Search Pricing") (2020 (accessed June 28, 2020)) URL https://rapidapi.com/contextualwebsearch/api/web-search/pricing

[75] Kakaly, twitter-feed (2019 (accessed July 18, 2020)). URL https://github.com/kakaly/twitterfeed

[76] Heroku, Heroku Pricing (n.d. (Accessed July 18, 2020)). URL https://www.heroku.com/pricing

[77] HuggingFace, roberta-base-openai-detector (2020 (Accessed December 28, 2020)). URL https://huggingface.co/robertabase-openai-detector

[78] L. Cui, D. Lee, Coaid: Covid-19 healthcare misinformation dataset (2020). arXiv:2006.00885.

[79] F. Hamborg, N. Meuschke, C. Breitinger, B. Gipp, news-please: A generic news crawler and extractor, in: M. Gaede, V. Trkulja, V. Petra (Eds.), Proceedings of the 15th International Symposium of Information Science, 2017, pp. 218–223.

[80] Tweepy, Tweepy: Twitter for Python! (2020 (Accessed July 18, 2020)). URL https://github.com/tweepy/tweepy

[81] ipapi, ipapi - IP Address Lookup and Geolocation API (2020 (Accessed November 16, 2020)). URL https://ipapi.co/

[82] E. C. Tandoc, Jr., D. Lim, R. Ling, Diffusion of disinformation: How social media users respond to fake news and why, Journalism 21 (3) (2020) 381–398. arXiv: https://doi.org/10.1177/1464884919868325, doi:10.1177/1464884919868325. URL https://doi.org/10.1177/1464884919868325

[83] T. Rasool, W. H. Butt, A. Shaukat, M. U. Akram, Multi-label fake news detection using multi-layered supervised learning, in: Proceedings of the 2019 11th International Conference on Computer and Automation Engineering, ICCAE 2019, Association for Computing Machinery, New York, NY, USA, 2019, p. 73–77. doi:10.1145/3313991.3314008. URL https://doi.org/10.1145/3313991.3314008

[84] J. Ratkiewicz, M. Conover, M. Meiss, B. Goncalves, A. Flammini, F. Menczer, Detecting and tracking political abuse in social media (2011). URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2850/3274

[85] G. Giotta, Ways of Seeing . . . What You Want: Flexible Visuality and Image Politics in the Post-Truth Era, in: Fake News: Understanding Media and Misinformation in the Digital Age, The MIT Press, 2020, pp. 29–44. arXiv: https://direct.mit.edu/book/chapter-pdf/271929/9780262357388_cak.pdf, doi:10.7551/mitpress/11807.003.0005. URL https://doi.org/10.7551/mitpress/11807.003.0005

[86] M. Gao, Z. Xiao, K. Karahalios, W.-T. Fu, To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles, Proc. ACM Hum.-Comput. Interact. 2 (CSCW) (Nov. 2018). doi: 10.1145/3274324. URL https://doi.org/10.1145/3274324

[87] I. Skurnik, C. Yoon, D. C. Park, N. Schwarz, How Warnings about False Claims Become Recommendations, Journal of Consumer Research 31 (4) (2005) 713–724. doi:10.1086/426605. URL https://doi.org/10.1086/426605

[88] L. Hasher, D. Goldstein, T. Toppino, Frequency and the conference of referential validity, Journal of Verbal Learning and Verbal Behavior 16 (1) (1977) 107 – 112. doi: https://doi.org/10.1016/S0022-5371(77)80012-1. URL http://www.sciencedirect.com/science/article/pii/S0022537177800121

[89] S. A. Hawkins, S. J. Hoch, Low-Involvement Learning: Memory without Evaluation, Journal of Consumer Research 19 (2) (1992) 212–225. arXiv: https://academic.oup.com/jcr/article-pdf/19/2/212/5438798/19-2-212.pdf, doi:10.1086/209297. URL https://doi.org/10.1086/209297

[90] I. M. Begg, A. Anas, S. Farinacci, Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth, Journal of Experimental Psychology: General 121 (4) (1992) 446–458. doi:10.1037/0096-3445.121.4.446.

[91] D. T. Gilbert, D. S. Krull, P. S. Malone, Unbelieving the unbelievable: Some problems in the rejection of false information, Journal of Personality and Social Psychology 59 (4) (1990) 601–613.

[92] D. Polage, Source Credibility and Belief in Fake News: I'll Believe You If You Agree with Me, in: Fake News: Understanding Media and Misinformation in the Digital Age, The MIT Press, 2020, pp. 235–244. arXiv: https://direct.mit.edu/book/chapter-pdf/271956/9780262357388_ceb.pdf, doi:10.7551/mitpress/11807.003.0025. URL https://doi.org/10.7551/mitpress/11807.003.0025