Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

Cybersecurity Misinformation Detection on Social Media: Case Studies on Phishing Reports and Zoom's Threat

Mohit Singhal, Nihal Kumarswamy, Shreyasi Kinhekar, Shirin Nilizadeh

The University of Texas at Arlington
(mohit.singhal, nihal.kumarswamy, shreyasi.kinhekar)@mavs.uta.edu, shirin.nilizadeh@uta.edu

Abstract

Prior work has extensively studied news, politics, and health misinformation. However, misinformation can also be about technological topics. While less controversial, such misinformation can severely impact companies' reputations and revenues and users' online experiences. Recently, social media has also been increasingly used as a novel source of knowledgebase for extracting timely and relevant security threats fed to threat intelligence systems for better performance. However, with possible campaigns spreading false security threats, these systems can become vulnerable to poisoning attacks. In this work, we proposed novel approaches for detecting misinformation about cybersecurity and privacy threats on social media, focusing on two topics with different types of misinformation: *phishing websites* and *Zoom's security & privacy threats*. We developed a framework for detecting inaccurate phishing claims on Twitter. Using this framework, we could label about 9% of URLs and 22% of phishing reports as misinformation. We also proposed another framework for detecting misinformation about Zoom's security and privacy threats on multiple platforms. Our classifiers showed great performance with more than 98% accuracy. Employing these classifiers on the posts from Facebook, Instagram, Reddit, and Twitter, we found that about 18%, 3%, 4%, and 3% of posts were misinformation, respectively. In addition, we studied the characteristics of misinformation posts, their authors, and their timelines, which helped us identify campaigns.

Introduction

Prior work has extensively studied misinformation related to news, politics, and health (Rashkin et al. 2017; Ruchansky, Seo, and Liu 2017; Love, Blumenberg, and Horowitz 2020; Singhal et al. 2023). Though misinformation can also be about technologies and tools that people use in their everyday life. While less controversial, such misinformation can severely impact companies' reputations and revenues and users' online experiences. Moreover, recently social media has been increasingly employed as a novel source of information for threat intelligence systems. For example, a recent study showed that 25% of vulnerabilities appear on social media before the National Vulnerability Database (Brettman 2020), and as a result, numerous threat intelligence tools, such as Spider-Foot (spi 2021) and IntelMQ (Int 2021), started collecting intelligence from social media platforms. Another recent study showed that Twitter phishing reports provide detailed information about phishing threats and include more sophisticated phishing threats that remain undetected by anti-phishing tools (Roy, Karanjit, and Nilizadeh 2021a). However, we argue that such tools are vulnerable to poisoning attacks because social media posts (1) can be posted by an adversary and (2) are weakly monitored, as detecting and removing such misinformation is currently not the priority of social media platforms.

On the other hand, novel approaches need to be employed to detect misinformation about technologies because compared to misinformation about news, politics, and health, their purpose and impact are different. One has roots in people's political and cultural views and beliefs, while the other might take advantage of people's lack of technical background, their beliefs about technology or company, or their fear of possible security threats.

No work has systematically studied the spread and characteristics of misinformation about technological topics. In this work, we study misinformation about cybersecurity and privacy threats on social media, focusing on: *phishing websites* and *Zoom's security & privacy threats*. We chose these two because the misinformation about each has different intentions, characteristics, and consequences. Through this work, we show that different approaches should be employed to detect these different types of misinformation.

Phishing is a type of social engineering attack through which attackers try to trick victims into disclosing their private and sensitive information (Dhamija, Tygar, and Hearst 2006). To inform other users, some social media users share reports of phishing websites (Roy, Karanjit, and Nilizadeh 2021a). However, false phishing reports can also circulate on these platforms. These false reports can decrease the websites' visits, especially if these websites are added to blocklists. They can also increase the false-positive rates of anti-phishing systems. Recent work showed the presence of false phishing reports (Roy, Karanjit, and Nilizadeh 2021a) on Twitter. However, in this work, we systematically analyze this threat and propose a detection algorithm that can, with high accuracy, identify false phishing reports in real-time.

With the surge in the use of video conferencing tools, such as Zoom (Koeze and Popper 2020) during the pandemic, came concerns about the company's handling of the security

Figure 1: Claim about Zoom

and privacy of its user base. Users discussed issues, such as Zoombombing (Redden 2020) and private Zoom meetings that can be available to the public (Harwell 2020). However, not all the discussions were accurate. For example, Figure 1 shows a claim that Zoom is a "Chinese spying tool." However, the author has not provided supporting evidence, and the claim is misleading (McCarthy 2020). In addition, the tweet claims that Zoom does not use any encryption service, which was not true when this tweet was posted (Barrett 2020). Discourses like this can impact the overall image of the company, as well as Zoom's (new) users' experiences. No other work has studied misinformation about Zoom or any other technology-related topic. Existing studies on fake news and vaccine misinformation also only focus on one social media platform, while we have investigated four platforms, including Twitter, Facebook, Instagram, and Reddit.

Both phishing and Zoom misinformation can be categorized as *fact-based* misinformation, as they are entities that have unique groundtruth values (Kumar and Shah 2018). For example, one can check if a website is phishing or if Zoom has a specific vulnerability. Therefore, we define misinformation based on criteria that help check these entities against their groundtruth values. We define a claim about a phishing website as a false report if the phishing link provided in the tweet refers to a benign website, and therefore the report is *false*. We define a post about Zoom's security and privacy threats as misinformation if it fails to provide supporting evidence and cannot be verified by cross-checking it with reputable sources or if the provided information is fully or partially in contrast with that of trusted sources. While both false phishing reports and Zoom claims can be intentional and, therefore, disinformation, in this paper, we do not differentiate between misinformation and disinformation, as we do not explore methods to detect intentions.

The examples and definitions of misinformation for these topics illustrate that not the same approach can be employed to detect these different types of misinformation. One naive solution to detect a false phishing report is to check if they appear on open anti-phishing sites, such as PhishTank or VirusTotal. However, these sites might not be up-to-date (Peng, Harris, and Sawa 2018; Kantchelian et al. 2015). To detect misinformation about Zoom security and privacy, not only does one need to have access to a knowledgebase that lists security and privacy threats of Zoom, but one also needs to consider the language used by the author, to discuss the matter, as unlike the phishing reports, the discussion over Zoom does not follow a certain structure or template, and users might be addressing the issue arbitrarily.

In this paper, we investigate three research questions: **RQ1:** How misinformation about phishing and Zoom are prevalent on social media? **RQ2:** How can such misinformation be detected? **RQ3:** What are the characteristics of misinformation posts, their authors, and campaigns?

To answer these research questions, we propose two frameworks for detecting and analyzing misinformation related to each topic. The first framework investigates the correctness of phishing reports on Twitter through a multi-step approach, which includes: (1) obtaining tweets about phishing, extracting unique URLs, (2) creating a groundtruth by regularly checking them via VirusTotal (Vir 2020), on PhishTank (phi 2020) and manually, and then (3) developing a classifier that can successfully detect false phishing reports in real-time. The second framework examines the correctness of posts about *Zoom*'s security and privacy threats by: (1) obtaining posts from Facebook, Instagram, Reddit, and Twitter, (2) using a ground truth dataset to identify the features that make misinformation posts distinguishable from accurate posts, and (3) using the features to build a classifier that detects misinformation. Thus, in this paper, we have the following contributions:

1. We showed social media users share false phishing reports and proposed a framework for detecting such posts.
2. We presented a new annotated groundtruth dataset for security & privacy issues regarding Zoom and identified misinformation features through quantitative and qualitative analysis. This dataset can be used as a benchmark by the community to build and test their proposed detection algorithms.
3. We developed classifiers that detect misinformation about Zoom's security and privacy on four social media platforms. Using these classifiers, we showed that such misinformation is prevalent, especially on some platforms, such as Facebook.
4. We characterized the detected misinformation posts, their authors, and possible campaigns.
5. We hope that this work increases the awareness of the community and social media platforms about the spread of misinformation about technological topics.

Related Work

Obtaining cybersecurity threats from social media. Recently, some works have proposed using social media, such as Twitter and Facebook, as the main source of identifying new vulnerabilities (Sabottke, Suciu, and Dumitraș 2015; Roy, Karanjit, and Nilizadeh 2021b). (Alves et al. 2021) introduced a Twitter streaming threat monitor that generates a continuously updated summary of the threat landscape related to a monitored infrastructure. (Okutan, Yang, and McConky 2017) integrated tweets with posts from the GDELT news service and Hackmageddon to detect new Defacement, DoS, and Malicious Email/URL threats. (Sapienza et al. 2017) introduced a system that leverages the communication of malicious actors on the dark web and the activity of security experts on Twitter to generate warnings of imminent or current cyber threats automatically.

Misinformation Detection in Social Media. A large body of work has tried detecting fake political news (Shu et al. 2017; Shu, Wang, and Liu 2019; Zhou, Wu, and Zafarani 2020), investigating various linguistic features (Hosseinimotlagh and Papalexakis 2018; Markowitz and Hancock 2014), as well as deep neural networks (Karimi et al. 2018; Karimi and Tang 2019; Wang et al. 2018). In recent months, scholars have analyzed misinformation related to COVID-19 (Brennen et al. 2020; Kouzy et al. 2020; Loomba et al. 2021; Singh et al. 2020). A recent work (Roy, Karanjit, and Nilizadeh 2021a) showed the presence of false phishing reports on Twitter. We analyze this threat more deeply and propose an algorithm for detecting false phishing reports. To the best of our knowledge, our work is the first study to find misinformation about security claims.

Framework for Detecting False Phishing Reports on Twitter

We propose a framework for detecting false phishing reports on Twitter. We conduct our study on Twitter because other works have shown the existence of phishing reports on this platform (Roy, Karanjit, and Nilizadeh 2021a). Figure 2 shows our proposed framework, which consists of (1) Data collection, (2) Groundtruth creation, (3) Feature selection, (4) Classification, and (5) Analysis of false phishing reports.

Data Collection

The data collection module is about collecting phishing reports on Twitter, and it consists of three steps: **(1) Collection of tweets:** Using the Twitter streaming API (Twi 2020), we collected 1% sample of daily tweets on Twitter that include the keyword *phishing*, from January 11, 2021, to April 11, 2021. We did not use Twitter V2 API because, at the time of data collection, it was not available (Twi 2021a). We considered retweets in our dataset. **(2) Pre-processing and filtering:** We filtered tweets to obtain those that were posted on the same day, as the Twitter API provides tweets posted over the last 7 days. **(3) URL extraction:** After manually inspecting a random sample of 100 tweets, we found that many tweets are *educational* and are about phishing threats in general or provide some stories about phishing events. We also found that the tweets with phishing reports usually include the obfuscated URL of the potential phishing website, i.e., using the following formats "hxxps[:]//xyz[.]com," or "hXXp:[//]xyz[dot]com." This is consistent with observation provided by (Roy, Karanjit, and Nilizadeh 2021a). Therefore, we employed regular expressions to retrieve the tweets with obfuscated URLs and obtain those tweets claiming that specific websites are phishing websites. We manually checked the obtained 100 tweets and found that the precision of this module is 1, and our final dataset only includes tweets with phishing claims. Finally, we extracted the URLs from these tweets to be validated.

Groundtruth Creation

For training the classifier, we created a groundtruth dataset of true and false phishing reports. This is not a trivial labeling task, requiring a longitudinal analysis of phishing reports. This module consists of the following steps: **(1) Daily analysis of URLs via VirusTotal:** We evaluated the URLs *daily* by passing them through VirusTotal API (Vir 2020). VirusTotal provides aggregated results for an URL obtained from 80 scan engines by third-party security vendors. Given a particular URL, the API returns the labels from all the vendors, and it shows the number of scan engines that detected the URL as *malicious* and *benign*. We analyzed the URLs daily because phishing campaigns can be short-lived, and some websites can go inactive or be re-sold. **(2) Delayed analysis of VirusTotal scores:** Prior work (Peng et al. 2019; Kantchelian et al. 2015; Zhu et al. 2020) has pointed out that VirusTotal is slow in updating its database, and there is a high probability that a website that was flagged as benign on the first day be flagged malicious later on. Therefore, we employed VirusTotal again on May 3, 2021, three weeks after the last day of data collection. **(3) Checking benign URLs on PhishTank:** Since anti-phishing engines can misclassify malicious URLs as benign, this module further passed all the URLs labeled as *benign* to PhishTank (phi 2020), a free community-based site for reporting phishing websites. **(4) Manual inspection of *benign* URLs:** Since it is also possible that the malicious URLs are not detected by VirusTotal and also have not been reported to the PhishTank, additionally, we manually checked the URLs that are labeled *benign* in the previous steps. Particularly, on a virtual machine, we checked whether the URL mimics a login page or prompted users to download something.

Groundtruth dataset and prevalence of false phishing reports: Table 1 shows our final groundtruth dataset. We found a union set of 11,472 users who posted 17,770 phishing reports. Among 10,578 unique URLs, we identified 9,603 as *malicious* and 975 as *benign*, corresponding to 13,875 and 3,895 tweets, respectively. Therefore, we can conclude that about 9% of all *obfuscated URLs*, and about 22% of tweets are *misinformation*. We found 11,472 and 148 unique users with true and false reports, respectively, with 124 users posting a mix of true and false reports.

Detecting False Phishing Report in Real-time

To identify phishing misinformation in real-time, we developed a classifier using our groundtruth dataset.

Textual and Contextual Feature Selection We used a combination of textual and contextual features. **Textual Features:** We extracted bi-grams and uni-grams from all the posts and considered the top 100 of them with the highest values of TF-IDF, which resulted in 64 uni-grams and 36 bi-grams. We tested our classifiers with 100, 500, 1000, 1500, and 2000 top n-grams and found that 100 provided the best results. **Contextual Features:** Along with the textual features obtained from the posts, we extracted a set of contextual features from the meta-data, which include: (1) *VirusTotal Scores:* A Higher VirusTotal score can be a great indication of a URL being malicious. However, in line with prior work (Peng et al. 2019; Zhu et al. 2020), we found that VirusTotal scores might not be accurate at the time of the report. Therefore, they cannot solely be trusted with detecting false phishing reports in real-time. We used Virus-

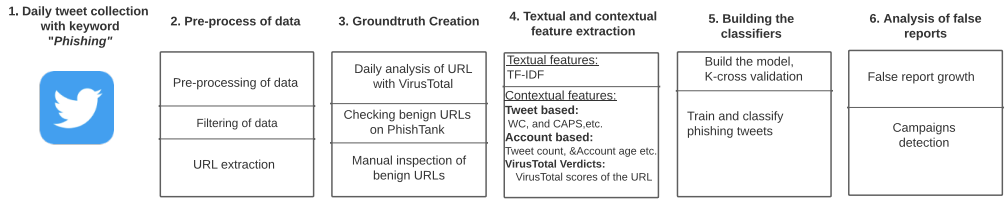


Figure 2: The framework developed for detecting false phishing reports on Twitter

# Tweets	Unique Users	Unique URLs	Malicious URLs	Benign URLs	Malicious Tweets	Benign Tweets (False claim)	Accounts with true claims	Accounts with false claims
17,770	11,472	10,578	9,603	975 (9%)	13,875	3,895 (22%)	11,200	148

Table 1: We found 22% of phishing reports to be false, as they were inaccurately reported as phishing websites.

Total scores, obtained at the time of Twitter data collection, as a feature in our classifier, and we also relied on additional features obtained from Twitter posts and authors, which might help distinguish a false report from a truthful report. (2) *Post-inspired features:* We used some features obtained from the posts, including *length of tweet*, and *the total number of capital words in a tweet*, which was used as a feature named *CAPS*. (3) *Features based on account characteristics:* Most prior work on phishing detection (Sun et al. 2016; Ma et al. 2009; Canali et al. 2011) uses the features captured from the URLs. Some works (Aggarwal, Rajadesingan, and Kumaraguru 2012; Chen et al. 2015), on detecting phishing URLs on social media, have also used features extracted from the social media posts, such as the author's features. We used the following account characteristics: *number of tweets*, *profile description length*, *account age*, *listed count*, *verified account*, *followers count*, and *has a profile image*. These features indicate if accounts are well-established, active, and anonymous, which might imply their trustworthiness (Morris et al. 2012; Zubiaga and Ji 2014).

Classifier We developed a binary classifier using our already labeled groundtruth dataset. We used the first 10 weeks of our groundtruth dataset for training and the remaining 3 weeks for testing. This setup evaluates the performance of our classifier on real-time data, i.e., phishing reports that are seen for the first time. We used 10,851 true claims posts, 2,434 false claims posts as our training set, and 3,024 & 1,461 true and false claim posts for our testing set, respectively. Before extracting the features, we performed preprocessing on our dataset, removed stop words, emojis, special characters (hashtags), and URLs, and performed stemming, i.e., cataloging related words together. These steps help minimize the sparsity of data (Da Silva, Hruschka, and Hruschka Jr 2014; Patwa et al. 2021), e.g., by not removing the # sign from #covid, the word *covid* would have been counted separately from every other covid word present in the posts. We vectorized our data using the TF-IDF. We found that a mix of uni-gram & bi-gram features provided a better result compared to just using either uni-gram or bi-gram or tri-gram. Since our groundtruth dataset for each platform was unbalanced, we employed several oversampling techniques, such as RandomOversampler, Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al. 2002), and Adaptive Synthetic Sampling (ADASYN) (He et al. 2008). We tested multiple classification algorithms, such as Random Forests, SVM, Naive-Bayes & K-nearest neighbor.

Classifier	Accuracy	F1 Score	Precision	Recall
Only VirusTotal	0.79 (+/- 0.01)	0.70 (+/- 0.02)	0.84 (+/- 0.02)	0.79 (+/- 0.02)
Only Contextual	0.92 (+/- 0.02)	0.93 (+/- 0.01)	0.93 (+/- 0.03)	0.92 (+/- 0.02)
All features	0.95 (+/- 0.02)	0.95 (+/- 0.03)	0.95 (+/- 0.02)	0.95 (+/- 0.01)

Table 2: The performance of classifiers

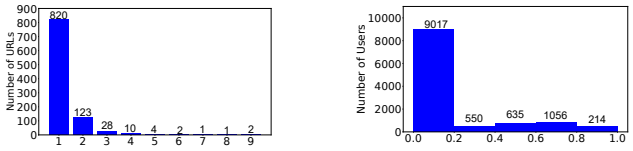
Classifier Performance Table 2 shows the results of our classifier. We found that only using VirusTotal scores, as the only feature, did not yield the best results. Adding other contextual features increased the performance, from 79% accuracy to 92%. We obtained the best results, i.e., 95% accuracy, precision, recall, and F1-score when we used all the textual and contextual. We found that the Random Forest classifier and SMOTE provided the best performance. We also examined feature importance in the trained Random Forest model to understand which of the features have a higher importance in classification tasks. The top 5 features and their scores are: *VirusTotal Score* (0.416), *Length of Tweet* (0.086), *Profile description length* (0.040), *phishing* (0.032), *possible threat* as a phrase (0.027).

Accounts and Campaign Characterization

Descriptive Statistics Table 3 statistically describes the user accounts with true and false claims. If a user had posted both true and false claims, we considered them in both sets. We compared the account characteristics of users with true claims vs. false claims. To compare features, such as *Followers, Friends, & Tweets*, we ran Mann-Whitney U tests as they do not follow a normal distribution. We could not reject the null hypothesis that users with *false* and *true* claims have the same distribution of followers counts, friends count, and tweet counts. We ran a chi-square test for the *verified* variable, and could not reject the null hypothesis.

Feature	Users with true claims/ Users with false claims Mean	Min	Max	Med.
Followers	6.2K/ 2K	0/ 0	12M/ 35K	302/ 426
Friends	1.4K/ 1.2K	0/ 0	275K/ 22K	416/ 458
Tweets	30K/ 60K	1/ 6	3.4M/ 2.2M	4.9K/ 5.4K
Verified	0.02/ 0.01	0/0	1/1	-

Table 3: Descriptive statistics of our final datasets.



(a) Tweet counts (b) False claim rate per user

Figure 3: Histogram of tweets and users for false claims.

Spread of Misinformation: Campaigns We constructed a network based on *following* relationships between all users in our false claim dataset using the Louvain Community Detection method (Blondel et al. 2008). In total, our network has 121 nodes and 618 edges. This is only a subset of the users, as we were not able to obtain the following 27 accounts. The average weighted degree is 5.1. The average path length from one randomly selected node to another is 2.47. In total, we were able to obtain 6 communities with 37, 34, 22, 22, 4, and 2 nodes. We found that 66 accounts posted the same false claims in their community. Also, the biggest community contributes to about 51% of the false reports.

Campaign Detection: Campaigns against Specific Websites. Figure 3a shows the histogram of tweet counts for all the websites in our false claim dataset. While most of the websites have been falsely reported in only one tweet, there are some campaigns where a phishing claim against a specific website has been tweeted multiple times. We identified 48 URLs, that were tweeted more than 3 times, by a total of 32 unique users. This suggests the existence of campaigns against specific websites.

Campaign Detection: Users with Many False Claims Figure 3b shows the histogram of users with false claims divided by their total number of phishing claims. We found that only a small number of users have posted many phishing claims. Interestingly, we found one user with 650 false claims and 3K true claims. About 78% of the users in our dataset, i.e., 8,942 users, have only true claims, while 29 users have only *false claims*. Also, 502 (4.37%) users have a false claim rate of around 0.5. Almost all of these users have an equal amount of tweets, i.e., one true claim and one false claim (310 users) or two true claims and two false claims (174 users). Users that have only posted *false claims* are suspicious and they might have deliberately sent these tweets.

Bots. We analyzed the false claim users using Botometer (Sayyadiharikandeh et al. 2020), a tool that gives a score to the accounts based on their social activity and other features. Using a threshold of 0.6, we found 2 bots and 139 real accounts, while 7 users changed their profiles to protect.

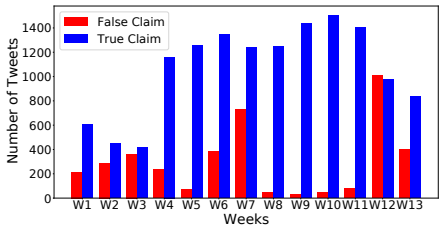


Figure 4: False vs. true phishing claims growth

Growth of phishing reports. Figure 4 shows the number of true and false phishing reports weekly from January 11, 2021 (the start of our data collection), to week 13, April 11, 2021 (the end of our data collection). As you can see, false reports do not follow a trend, and in some weeks, there is a huge growth in their number, e.g., week 7 and week 12, while in some other weeks, there is a huge drop in their number, e.g., in week 5 or week 8. In all weeks, however, except week 12, the number of true phishing reports is higher than that of the false reports. The sharp increase in the number of false reports in week 12 is in-line with the real-world event when the COVID-19 Delta variant became the dominant variant and the number of phishing scams related to COVID-19 increased (Larson 2021).

Discussion. Our results show that not all phishing reports are reliable and there is a need for mechanisms to validate their correctness. Our proposed framework in Figure 2 can be employed by social media platforms to detect and remove false phishing reports.

Framework for Detecting Misinformation about Zoom's Security and Privacy Threats

In this section, we focus on social media discussions around Zoom's security and privacy threats and propose a framework for detecting misinformation regarding it on Facebook, Instagram, Twitter, and Reddit. Analyzing public data from multiple social media platforms can also help us to investigate how misinformation is circulated differently on these platforms. To detect misinformation posts on each social media platform, we developed a binary classifier specific to that platform. Figure 5 shows our proposed framework: (1) Data collection, (2) Groundtruth and codebook creation, (3) Feature selection, (4) Training and testing classifiers, and (5) Detecting the misinformation in each platform.

Platform	2019	2020	After filtering
Instagram	42,639	422,874	6,885
Facebook	167,718	4,537,280	74,590
Reddit	21,250	134,866	9,167
Twitter	45,178	1,011,022	870,852

Table 4: The number of posts in each platform

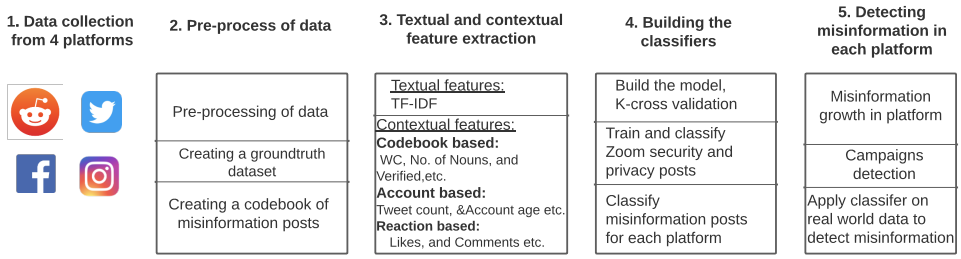


Figure 5: The framework developed for detecting Zoom security & privacy misinformation.

No.	Post	Label	Reason
1	Security researchers have called Zoom "a privacy disaster" and "fundamentally corrupt" as allegations of the company mishandling user data snowball #Data #Breach #Zoom https://t.co/r3NcjsmuAB	Zoom's security & privacy	Satisfies all the criteria and link goes to "Guardian" website
2	@XXX CEO @XXX: With the popularity of #Zoom, some security concerns have come to light- No end-to-end encryption for call- Sale of user data and analytics without disclosing or proper authorization. #UpskillGang #MilimaCyberAwareness @XXX' [Tweeted on April 24th, 2020]	Misinformation	Fails to provide evidence and that no end-to-end encryption is false ((Barrett 2020)) and that it sells user data ((Yuan 2020))
3	[#coronavirus] Japanese fashion brand now sells T-shirts for #Zoom mtg. Change the color and design but basically only simple green T-shirts. Using the technology of virtual back ground and change as you like See below news. Seems to be nice! https://t.co/8lTMxtzKZb	Irrelevant	Not about cybersecurity

Table 5: Examples of posts and assigned labels.

Platform	Zoom security	Misinformation	Irrelevant
Instagram	545	15	2,740
Facebook	560	42	2,734
Reddit	1,045	16	2,234
Twitter	1,865	36	1,468

Table 6: The size of groundtruth datasets per platform.

Data Collection

In order to collect data from Facebook, Instagram, Reddit, and Twitter, we used the "posts/search" endpoint of the CrowdTangle API (Team 2020). The CrowdTangle API provides about 2% of all public Facebook groups and pages, 2M+ public Instagram pages, and about 20K+ of most active sub-reddits (Fraser 2021). We also collected Twitter data using the Twitter V2 Archive Search endpoint (Twi 2021b), which allows us to search access public Tweets from the complete archive dating back to the first Tweet in March 2006. We collected English posts sent from June 1, 2019, to Nov. 30, 2020, to examine if users were discussing security and privacy issues of Zoom before the pandemic, and how the discussions changed when the pandemic started. We initially obtained posts that include the keyword *Zoom*. We considered retweets in our dataset.

Pre-processing and filtering: Table 4 shows the data collected from the four platforms. Since we collected the posts that included the keyword "Zoom," our dataset contained many posts not talking about security and privacy. To find additional keywords, we used the snowball sampling technique (Goodman 1961). We started by using a couple of seed keywords, including Zoom, Security, and Privacy. We then extracted posts from our dataset for the month of March for each respective platform. Using the seed keywords, we iteratively identified potential keywords that frequently co-occur with the seeds, adding them to our seed list only after manually ensuring they are closely related to our topic. After saturation was reached, we manually combined keywords into composites. In total, we identified 18 such phrases, all starting with the word *Zoom* and then the following words: *Malware, Phishing, Virus, Security, Exploit, Hijacking, Bug, Hackers, Privacy, Backdoor, Hacked, Security Bug, Windows, Passwords, Windows Steal, Zoombombing, and Data*. We then used our new expanded keyword list, to filter out the posts. The last column of Table 4 shows the final dataset that was obtained after our filtering.

Groundtruth Dataset

Groundtruth Creation For training the classifiers, we first manually labeled a subset of the posts on each platform to create a groundtruth dataset. Creating a groundtruth is not a trivial task because we need to verify the correctness of claims and discussions. We used the following three criteria to label the posts: (a) The post is talking about "Zoom," (b) The post is talking about Zoom's security or privacy, and (c) The post is either providing some evidence, i.e., links,

videos, etc., from reputable blogs or articles that are verifiable, or not providing supporting evidence, but we could verify the claims by cross-checking them with the reputable sources. For that, we checked the post context and ran a Google search to determine if the post context is already addressed by the company or reputable sources and if the claim can be validated. Using these criteria, we defined three labels: *(1) Zoom's security and privacy:* if a post satisfies all of the above-mentioned criteria, *(2) Misinformation:* if the third criterion is not satisfied, and *(3) Irrelevant:* if it fails to satisfy either the first or second criteria. Some examples of the posts that were labeled are shown in Table 5.

Annotation Process To hand-label the posts, two authors annotated 13,200 posts (3,300 randomly chosen posts from each of Twitter, Instagram, Facebook & Reddit). For inconsistent results, coders discussed how to resolve disagreements. To assess the inter-coder reliability, we performed a Cohen-Kappa test (Schuster 2004). The inter-rater agreement measured with the Kappa score was 0.972, which shows almost perfect agreement. Table 6 shows the groundtruth dataset that was obtained after the annotation. We can observe that there are a significant number of *irrelevant* posts in our dataset, however, we do find evidence of misinformation in our dataset. We found 23 instances of users inviting other users to "Zoombomb" their classes or meetings.

Qualitative Analysis of Misinformation Posts

We qualitatively analyzed the misinformation posts to understand their types, targets, and properties. These properties then can be used as features for classification. We created a hierarchical codebook of misinformation about Zoom's security and privacy threats, applying the open coding process (Glaser and Strauss 2017). Following this process, one of the authors coded the misinformation posts identified in the previous subsection until no new categories emerged. To improve the quality of the categories, we used an iterative process (Corbin and Strauss 1990) so that new categories were added or existing ones were reorganized. To create the codebook, we followed certain guidelines: (1) Read through the posts, and identify themes and sub-themes; (2) While creating the categories, identify the motive and meaning of the post; and (3) Consider various features that can help in the identification of categories.

Figure 6 shows the hierarchical structure of the codebook. We discovered 4 main topics: (i) *Sources*: Posts that provide misleading sources videos, URLs, or invalid links to other sites, (ii) *Structural*: Posts containing irregularities in the content like misspellings or written in capitalization, (iii) *Network*: Relates to the reach or perceived audience of the author, and (iv) *Post Type*: subdivided into 4 categories talking about security or privacy and text that can be misleading or accusing and contain logic flaws, biased authors, or propagating conspiracy. Table 7 gives a high-level overview of the description of each of these classes. These classes describe the data, and a post might fit multiple classes. For example, a post can be misleading suggesting that Zoom has no encryption and simultaneously accusing the company to

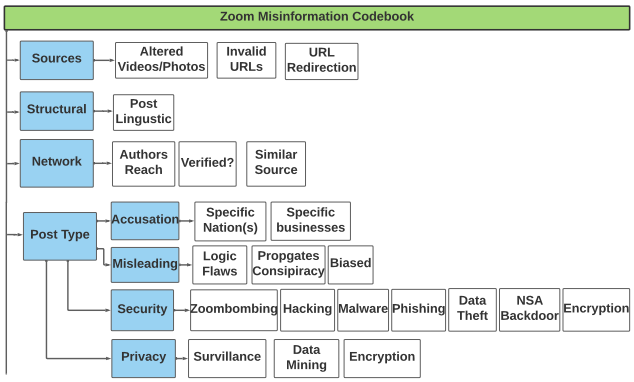


Figure 6: Zoom misinformation hierarchical codebook

be spyware from the Chinese Communist Party.

After saturation, two authors coded the 109 posts labeled as misinformation. To find the agreement score, we gave a value of 1, if two had a perfect agreement, otherwise, we divided the number of labels by the number of possible values. Using this methodology, we found a substantial agreement of 72.3%. The distribution of posts among different classes are: Sources (12), Structural (20), Network (11), Accusation (45), Misleading (20), Security (102), and Privacy (67). Note that one post could be assigned to multiple classes.

Detecting Zoom's Security and Privacy Threats

Initially, we tried to use a binary classifier to differentiate misinformation from all other posts. However, this yielded mediocre results. We then employed a two-step approach, where we first built a binary classifier to detect posts related to Zoom's security and privacy, and then built another classifier to detect misinformation among them. For the first classifier, we used the groundtruth dataset, where we used the misinformation posts and Zoom's security and privacy posts. We built one classifier for each of the four platforms as each platform gives a different style of data, e.g., Twitter allows up to 280 characters while there is no constraint on that of Facebook posts. To build our supervised classifiers, we found that a mix of uni-gram & bi-gram features provide a better result compared to uni-gram, bi-gram, or tri-gram. Before extracting the features, we performed pre-processing on our dataset and removed stop words, emojis, special characters (e.g., hashtags), and URLs. We vectorized our data using the TF-IDF. Since our groundtruth dataset for each platform was unbalanced (see Table 6), we employed several oversampling techniques. We also tested multiple classification algorithms. To evaluate our classifiers, we used k-cross validation, where $k = 3$. From our analysis we found that RandomOversampler was the best for Instagram, however, SMOTE provides better results for Facebook, Twitter, and Reddit. Table 8 shows the classification performance of the Random Forrest classifier as it provided the best accuracy across all four platforms. After developing classifiers detecting posts related to Zoom security and privacy, we ran the classifier on the whole datasets of all the platforms. Table 9 shows the percentages of security and privacy-related posts

Class	Sub-Class	Description
Sources	-	Providing altered videos or photos that are not in the context to create confusion; posting URLs that are invalid or are redirected to a third party site.
Structural	-	Post has all CAPS headline and content, and misspelling in the content.
Network	-	Has a large audience, is verified by the platform. Two or more sources show the same news with the same context over time.
Post Type	Accusation	Accusing various countries and/or businesses of wrongdoing without relevant evidence.
	Misleading	Misleading the audience, promoting and solidifying a myth that rejects accepted narrative, is aligned towards one way of thinking and draws conclusions based on a limited number of facts.
	Security	Post is about fake Zoombombing attacks, sponsors notion that using Zoom leads to hacking, data theft, leads to the backdoor for NSA, is malware and suggests no encryption used for chats and using Zoom leads to phishing attacks.
	Privacy	Post suggests that users are being watched by government, promotes that user data is being mined by other companies, and sponsors the notion that since no encryption, anyone can read your chats.

Table 7: Description of categories identified by qualitatively analyzing the misinformation posts

Platform	Accuracy	F1-Score	Precision	Recall
Instagram	0.94 (+/- 0.02)	0.93 (+/- 0.02)	0.93 (+/- 0.04)	0.94 (+/- 0.01)
Facebook	0.91 (+/- 0.02)	0.91 (+/- 0.02)	0.91 (+/- 0.02)	0.91 (+/- 0.04)
Reddit	0.93 (+/- 0.01)	0.93 (+/- 0.02)	0.93 (+/- 0.02)	0.93 (+/- 0.03)
Twitter	0.81 (+/- 0.02)	0.81 (+/- 0.01)	0.81 (+/- 0.03)	0.81 (+/- 0.01)

Table 8: Performance of classifiers detecting posts about Zoom security and privacy.

Platform	Zoom's security & privacy	Irrelevant
Instagram	551 (8%)	3,034
Facebook	11,400 (15%)	59,890
Reddit	627 (7%)	5,240
Twitter	505,418 (58%)	365,434

Table 9: Percentage of Zoom's security and privacy posts.

in our Instagram, Facebook, Reddit, and Twitter datasets, which are 8%, 15%, 7%, and 58%, respectively.

Detecting Misinformation about Zoom's Security and Privacy Threats

We trained another binary classifier for each platform to detect misinformation among posts relevant to Zoom security.

Feature Selection We used a combination of textual and contextual features. **Textual Features:** For each platform, we extracted bi-grams and uni-grams from all the posts and considered the top 100 of them with the highest values of TF-IDF, which resulted in 37 uni-grams and 63 bi-grams. We tested our classifiers with 100, 500, 1000, 1500, and 2000 top n-grams. While 500 provided the best results for Instagram, 100 provided the best results for other platforms. **Contextual Features:** We extracted a set of contextual features from the meta-data, including (1) *Features inspired by the qualitative analysis:* Creating the codebook, we found some features being more apparent in misinformation. For example, in terms of network structure, they tend to have a large audience and are verified, or in terms of sources, they tend to provide altered videos or photos. Therefore, we used the following features: *word counts*, *noun counts*, *pronouns counts*, the total number of capital words in a tweet, i.e., *CAPS*, *misspelled words count*, *verified account*, *followers count*, *has a photo/video* and *has a URL*. (2) *Reaction-inspired features:* Posts can get reactions, such as *likes*, *retweets/shares*, *comments*, etc. We used the following features: for Instagram, *likes count*, for Facebook, the number of *likes, comments*, and *shares*, for Reddit, the number of *likes*, and *comments*. For Twitter, we did not find the distribution of a number of *likes* and *retweets* being statistically different among the two classes, therefore we did not use them. (3) *Features based on account characteristics* including: *tweets count*, *profile description length*, *account age*, *listed count*, and *has a profile image*.

Classifiers Testing several oversampling techniques, we found that RandomOverSampler provides the best results for Instagram and Reddit classifiers, and SMOTE provides the best results for Facebook and Twitter classifiers. Also, we found that out of the five machine learning algorithms, Random Forest provides the best accuracy across the four platforms. We compared the results of the different algorithms after performing hyperparameter tuning, which conducts an exhaustive search over the parameters to find the best combination of parameters. Table 10 shows all classifiers, using k-cross validation ($k = 3$), have great performances.

We examined feature importance in the trained Random Forest model to understand which of the features have a higher importance in the classification tasks. The top 5 features and their scores for each model are: **Facebook:** no. of all CAPS words (0.083), word count (0.081), *company* (0.052), *behind* (0.046), *hey* (0.045). **Reddit:** *company* (0.116), *China* (0.087), no. of likes (0.066), word count (0.060), *security* (0.050). **Instagram:** *Zoom* (0.084), No. of all CAPS words (0.034), word count (0.033), *security* (0.029), *away* (0.029). **Twitter:** Has photo/video (0.072),

Platform	Accuracy	F1 Score	Precision	Recall
Instagram	0.98 (+/- 0.01)	0.98 (+/- 0.02)	0.98 (+/- 0.01)	0.98 (+/- 0.02)
Facebook	0.99 (+/- 0.00)	0.99 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.01)
Reddit	0.99 (+/- 0.01)	0.99 (+/- 0.01)	0.99 (+/- 0.00)	0.99 (+/- 0.01)
Twitter	0.98 (+/- 0.01)	0.98 (+/- 0.02)	0.98 (+/- 0.02)	0.98 (+/- 0.01)

Table 10: The performance of classifiers detecting misinformation about Zoom

	Users with true claims/ Users with misinformation			
Feature	Mean	Min	Max	Med.
Followers	9.3K/ 10K	0/ 0	58M/ 8.4M	520/675
Friends	1.5K/ 1.7K	0/ 0	1M/ 280K	678/ 735
Tweets	39K/ 10K	1/ 1	4.1M/ 3M	42K/ 10K
Verified	0.03/ 0.03	0/0	1/1	-

Table 11: Descriptive statistics of our final Twitter dataset.

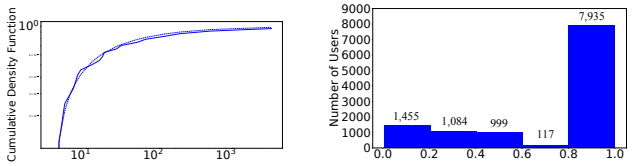
URL in Tweet (0.070), account age (0.049), *say* (0.043), has a profile image (0.042). The top features for each platform span a variety of feature categories, including textual features such as n-grams, and contextual features, consisting of reaction-, and account-inspired features.

Prevalence of misinformation about Zoom Finally, we employed our trained classifiers on the posts that are related to security and privacy to detect those that are misinformation. We found that overall about 3%, 18%, 4%, and 3% of posts on Instagram, Facebook, Reddit, and Twitter are *misinformation*, respectively. We verified if our classifiers show consistent performance by manually labeling 200 (100 misinformation, 100 Zoom S&P) posts on Twitter, 200 (100 misinformation, 100 Zoom S&P) posts on Facebook, 116 (16 misinformation, 100 Zoom S&P) posts on Instagram, and 125 (25 misinformation, 100 Zoom S&P) from Reddit, and computing the accuracy and F1-score on these testing tests. Two authors coded the random set, and for disagreements, they discussed having a final label. The Cohen-Kappa score was 0.784, showing substantial agreement. We found that about 93% and 92% accuracy and F1-score in the case of Instagram, 94% and 92% in the case of Reddit, 93% and 93% for Facebook, and 90% and 90% for Twitter.

Accounts and Campaign Characterization

We studied characteristics of the accounts that posted misinformation, mostly focusing on the Twitter dataset, because CrowdTangle does not provide meta-data about the authors.

Descriptive Statistics: Table 11 statistically describes the account characteristics of users with true and false claims. The number of unique users with true and false claims was about 220K and 11K, respectively. If a user had posted both true and false claims, we considered them in both sets. To compare features, such as *Followers, Friends, & Tweets*, we ran Mann-Whitney U tests as they do not follow a normal distribution. We could reject the null hypothesis that users



(a) CDF of #retweets (b) Misinformation rate per user

Figure 7: CDF & histogram of misinformation claims.

with *false* and *true* claims have the same distribution of followers count, and therefore, we can argue that on average, users with *true* claims have a lower number of followers than users with *misinformation* claims ($Med_{true} = 520$ vs. $Med_{mis} = 675$, $p < 0.0001$). We could not reject the null hypothesis for friends count & tweet count. We ran chi-square test for the *verified* variable, and found that accounts who post *misinformation* are more likely to be verified ($M_{mis} = 0.03$ vs. $M_{true} = 0.03$) ($X^2 = 17.27$, $p < 0.0001$). This is interesting and also in line with reports that *verified* users are sharing misinformation at an all-time high (Cohen 2021; Wang, Pang, and Pavlou 2018).

Spread of Misinformation: Campaigns We constructed a network based on *following* relationships between 3,564 unique users in our misinformation dataset, in the month of April 2020, as it has the highest number of misinformation, using the Louvain Community Detection method (Blondel et al. 2008). In total, our network has 2,975 nodes and 22,934 edges. We could not collect the following list for 80 of the users because they made their profiles protected. We also found that 500 users were not connected to any of the users. The average weighted degree is 7.70. The average path length from one randomly selected node to another is 3.63. These values show that these nodes are very connected to each other. In total, we were able to obtain 35 communities, including some large ones with 671, 618, 585, 499, 194, 134, and 107 nodes. The biggest community has 22.55% of all the nodes, while 13 communities have only 2 nodes.

Spread of Misinformation: Re-Tweet Count: Figure 7a shows the CDF of retweet counts for all the tweets in our misinformation dataset. While most of the misinformation was tweeted only once, we found some campaigns. We found one misinformation tweet that was retweeted 855 times. This retweet was sent by 854 unique users. The topic of the tweet was accusing Zoom of being a Chinese Communist Party malware and that they are using it for surveillance.

Campaign Detection: Users with Many False Claims: Figure 7b shows the histogram of users with misinformation claims divided by their total number of Zoom tweets. We found that a majority number of users have posted many Zoom misinformation claims. About 68% of the users in our dataset, i.e., 7,935 users, have only *misinformation claims*. Also, 955 (8.24%) users have a false claim rate of around 0.5. Almost all of these users have an equal amount of tweets, i.e., one true claim and one false claim (842 users) or two true claims and two false claims (113 users).

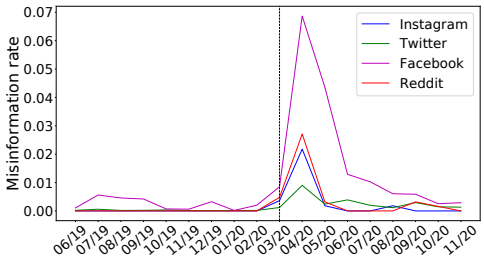


Figure 8: Misinformation growth rate.

Bots. We analyzed all the misinformation users using Botometer (Sayyadiharikandeh et al. 2020), and threshold of 0.6. We found 208 bots and 11,274 real accounts, while 107 users had changed their profiles to be protected.

Growth of Misinformation Figure 8 shows the percentage of misinformation over time on the four social media platforms. The vertical dotted black line represents the time when multiple states in the US went into COVID-19 lockdown (cov 2020). As you can see, on Facebook, there are some misinformation about Zoom since 2019. We manually inspected these posts and found that in the month of July 2019, users were discussing *Zoom hit by DoS*. However, while Zoom indeed had a new zero-day vulnerability that could be used to employ a DoS attack, no actual attack happened (Seals 2019). We also see a sudden spike, in the number of misinformation posts on Facebook around February 2020, and then subsequent spikes in Instagram, Twitter, and Reddit after March 2020. Our analysis revealed that users were claiming that Zoom is malware, a tool by the Chinese Communist party to spy on people, etc., however, these claims have been refuted (Barrett 2020; Yuan 2020). The plot shows a higher percentage of misinformation posted on Facebook when compared to other platforms, and at its peak, it gets to 7% of posts being misinformation.

Discussion. We showed that misinformation about Zoom security and privacy was spread on all the social media platforms. We argue that there is a need for mechanisms validating information about technological topics. Our proposed framework in Figure 5 is the first step toward such a goal.

Ethical Considerations and Broader Impact

We analyzed publicly available data, provided by Twitter Streaming API and Crowdtangle API. We also follow standard ethical guidelines (Rivers and Lewis 2014), not making any attempts to track users across sites or deanonymize them. We believe that our results show that misinformation about cybersecurity and privacy exists, and we hope that the community can further investigate its impact on the user and can research solutions to tackle this challenge.

Limitations and Future Work

The analysis on phishing misinformation gives a lower bound of the *misinformation* on Twitter because we had access to a 1% sample of Twitter data. Similarly, in our second study, the size of our datasets was restricted by *CrowdTangle*. Also, not having access to the followers and friends of users on other social media platforms, we could only detect and analyze possible campaigns on Twitter. In the future, we would investigate the diffusion of cybersecurity and privacy misinformation and examine if they are different from other types of misinformation. We could also explore using semi-supervised techniques instead of supervised learning.

Conclusion

In this work, we proposed two frameworks for detecting misinformation about cybersecurity and privacy threats on social media, focusing on two topics with different types of misinformation: *phishing websites* and *Zoom's security & privacy threats*. We examined the correctness of Twitter reports posted by users about websites being phishing. In total, we found that about 9% of all obfuscated URLs and about 22% of tweets about phishing websites are *misinformation*. Second, using a set of textual and contextual features, we built supervised classifiers to identify posts discussing the security and privacy of Zoom, and to detect misinformation in our whole dataset. Our classifiers showed great performance across all four platforms. We found about 3%, 18%, 4%, and 3% of posts on Instagram, Facebook, Reddit, and Twitter, as misinformation, respectively. Our results show that misinformation about cybersecurity and privacy is present on social media, and the community needs to further study its impact on end-users and threat intelligence tools.

Acknowledgments

This work is supported by NSF under CNS-1932574 and a Comcast Innovation Fund.

References

2020. COVID-19 lockdowns. https://en.wikipedia.org/wiki/COVID-19_lockdowns, (accessed on 10/12/21).

2020. PhishTank. https://www.phishtank.com/faq.php, (accessed on 12/17/20).

2020. Twitter Developer. https://developer.twitter.com/en, (accessed on 06/13/20).

2020. VirusTotal API. https://developers.virustotal.com/reference, (accessed on 01/13/21).

2021. IntelMQ. http://github.com/certtools/intelmq/, (accessed on 08/13/21).

2021. SpiderFoot, Open Source Intelligence Automation. http://spiderfoot.net/, (accessed on 08/13/21).

2021a. Twitter API v2 support. https://developer.twitter.com/en/support/twitter-api/v2, (accessed on 05/15/22).

2021b. Twitter V2 API. https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction, (accessed on 05/01/22).

Aggarwal, A.; Rajadesingan, A.; and Kumaraguru, P. 2012. PhishAri: Automatic realtime phishing detection on twitter. In *2012 eCrime Researchers Summit*, 1–12. IEEE.

Alves, F.; Bettini, A.; Ferreira, P. M.; and Bessani, A. 2021. Processing tweets for cybersecurity threat awareness. *Information Systems*, 95: 101586.

Barrett, B. 2020. Zoom Finally Has End-to-End Encryption. Here's How to Use It. https://www.wired.com/story/how-to-enable-zoom-encryption/, (accessed on 02/14/21).

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008.

Brennen, J. S.; Simon, F.; Howard, P. N.; and Nielsen, R. K. 2020. Types, sources, and claims of COVID-19 misinformation. *Reuters Institute*, 7: 3–1.

Brettman, A. 2020. Software Flaws Sometimes First Reported on Social Media. https://www.pnnl.gov/news-media/software-flaws-sometimes-first-reported-social-media, (accessed on 12/17/20).

Canali, D.; Cova, M.; Vigna, G.; and Kruegel, C. 2011. Prophiler: a fast filter for the large-scale detection of malicious web pages. In *Proceedings of the 20th international conference on World Wide Web*, 197–206.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

Chen, C.; Zhang, J.; Chen, X.; Xiang, Y.; and Zhou, W. 2015. 6 million spam tweets: A large ground truth for timely Twitter spam detection. In *2015 IEEE international conference on communications (ICC)*, 7065–7070. IEEE.

Cohen, J. 2021. Verified Twitter Users Shared an All-Time-High Amount of Fake News in 2020. https://www.pcmag.com/news/verified-twitter-users-shared-an-all-time-high-amount-of-fake-news-in-2020, (accessed on 05/14/22).

Corbin, J. M.; and Strauss, A. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1): 3–21.

Da Silva, N. F.; Hruschka, E. R.; and Hruschka Jr, E. R. 2014. Tweet sentiment analysis with classifier ensembles. *Decision support systems*, 66: 170–179.

Dhamija, R.; Tygar, J. D.; and Hearst, M. 2006. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 581–590.

Fraser, L. 2021. What data is CrowdTangle tracking? https://help.crowdtangle.com/en/articles/1140930-what-data-is-crowdtangle-tracking, (accessed on 01/17/21).

Glaser, B. G.; and Strauss, A. L. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Goodman, L. A. 1961. Snowball sampling. *The annals of mathematical statistics*, 148–170.

Harwell, D. 2020. Thousands of Zoom video calls left exposed on open Web. https://www.washingtonpost.com/technology/2020/04/03/thousands-zoom-video-calls-left-exposed-open-web/, (accessed on 10/09/21).

He, H.; Bai, Y.; Garcia, E. A.; and Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, 1322–1328. IEEE.

Hosseinimotlagh, S.; and Papalexakis, E. E. 2018. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. In *Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*.

Kantchelian, A.; Tschantz, M. C.; Afroz, S.; Miller, B.; Shankar, V.; Bachwani, R.; Joseph, A. D.; and Tygar, J. D. 2015. Better malware ground truth: Techniques for weighting anti-virus vendor labels. In *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*, 45–56.

Karimi, H.; Roy, P.; Saba-Sadiya, S.; and Tang, J. 2018. Multi-source multi-class fake news detection. In *Proceedings of the 27th international conference on computational linguistics*, 1546–1557.

Karimi, H.; and Tang, J. 2019. Learning Hierarchical Discourse-level Structure for Fake News Detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3432–3442.

Koeze, E.; and Popper, N. 2020. The Virus Changed the Way We Internet. https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html, (accessed on 12/10/20).

Kouzy, R.; Abi Jaoude, J.; Kraitem, A.; El Alam, M. B.; Karam, B.; Adib, E.; Zarka, J.; Traboulsi, C.; Akl, E. W.; and Baddour, K. 2020. Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus*, 12(3).

Kumar, S.; and Shah, N. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*.

Larson, S. 2021. As Delta Variant Spreads, COVID-19 Themes Make Resurgence In Email Threats. https://www.proofpoint.com/us/blog/threat-insight/delta-variant-spreads-covid-19-themes-make-resurgence-email-threats, (accessed on 05/04/22).

Loomba, S.; de Figueiredo, A.; Piatek, S. J.; de Graaf, K.; and Larson, H. J. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour*, 5(3): 337–348.

Love, J. S.; Blumenberg, A.; and Horowitz, Z. 2020. The parallel pandemic: Medical misinformation and COVID-19: Primum non nocere. *Journal of general internal medicine*, 35: 2435–2436.

Ma, J.; Saul, L. K.; Savage, S.; and Voelker, G. M. 2009. Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1245–1254.

Markowitz, D. M.; and Hancock, J. T. 2014. Linguistic traces of a scientific fraud: The case of Diederik Stapel. *PloS one*, 9(8): e105937.

McCarthy, B. 2020. PolitiFact. https://www.politifact.com/factchecks/2020/apr/07/charlie-kirk/china-spying-you-through-zoom-charlie-kirk-oversta/, (accessed on 04/14/21).

Morris, M. R.; Counts, S.; Roseway, A.; Hoff, A.; and Schwarz, J. 2012. Tweeting is believing? Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 441–450.

Okutan, A.; Yang, S. J.; and McConky, K. 2017. Predicting cyber attacks with bayesian networks using unconventional signals. In *Proceedings of the 12th Annual Conference on Cyber and Information Security Research*, 1–4.

Patwa, P.; Sharma, S.; Pykl, S.; Guptha, V.; Kumari, G.; Akhtar, M. S.; Ekbal, A.; Das, A.; and Chakraborty, T. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, 21–29. Springer.

Peng, P.; Yang, L.; Song, L.; and Wang, G. 2019. Opening the blackbox of virustotal: Analyzing online phishing scan engines. In *Proceedings of the Internet Measurement Conference*, 478–485.

Peng, T.; Harris, I.; and Sawa, Y. 2018. Detecting phishing attacks using natural language processing and machine learning. In *2018 ieee 12th international conference on semantic computing (icsc)*, 300–301. IEEE.

Rashkin, H.; Choi, E.; Jang, J. Y.; Volkova, S.; and Choi, Y. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2931–2937.

Redden, E. 2020. 'Zoombombing' Attacks Disrupt Classes. https://www.insidehighered.com/news/2020/03/26/zoombombers-disrupt-online-classes-racist-pornographic-content, (accessed on 02/10/21).

Rivers, C. M.; and Lewis, B. L. 2014. Ethical research standards in a world of big data. *F1000Research*, 3.

Roy, S. S.; Karanjit, U.; and Nilizadeh, S. 2021a. Evaluating the effectiveness of Phishing Reports on Twitter. In *2021 APWG Symposium on Electronic Crime Research (eCrime)*, 1–13. IEEE.

Roy, S. S.; Karanjit, U.; and Nilizadeh, S. 2021b. What Remains Uncaught?: Characterizing Sparsely Detected Malicious URLs on Twitter. In *MadWeb NDSS*, 400–412.

Ruchansky, N.; Seo, S.; and Liu, Y. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 797–806.

Sabottke, C.; Suciu, O.; and Dumitraș, T. 2015. Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting {Real-World} Exploits. In *24th USENIX Security Symposium (USENIX Security 15)*, 1041–1056.

Sapienza, A.; Bessi, A.; Damodaran, S.; Shakarian, P.; Lerman, K.; and Ferrara, E. 2017. Early warnings of cyber threats in online discussions. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 667–674. IEEE.

Sayyadiharikandeh, M.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2020. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2725–2732.

Schuster, C. 2004. A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, 64(2): 243–253.

Seals, T. 2019. Zoom Zero-Day Bug Opens Mac Users to Webcam Hijacking. https://threatpost.com/zoom-zero-day-mac-webcam-hijacking/146317/, (accessed on 10/11/21).

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1): 22–36.

Shu, K.; Wang, S.; and Liu, H. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, 312–320.

Singh, L.; Bansal, S.; Bode, L.; Budak, C.; Chi, G.; Kawintiranon, K.; Padden, C.; Vanarsdall, R.; Vraga, E.; and Wang, Y. 2020. A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv preprint arXiv:2003.13907*.

Singhal, M.; Ling, C.; Paudel, P.; Thota, P.; Kumarswamy, N.; Stringhini, G.; and Nilizadeh, S. 2023. SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice. arXiv:2206.14855.

Sun, B.; Akiyama, M.; Yagi, T.; Hatada, M.; and Mori, T. 2016. Automating URL blacklist generation with similarity search approach. *IEICE TRANSACTIONS on Information and Systems*, 99(4): 873–882.

Team, C. 2020. CrowdTangle. Facebook, Menlo Park, California, United States. https://www.crowdtangle.com/ (accessed on 01/15/21).

Wang, S. A.; Pang, M.-S.; and Pavlou, P. A. 2018. 'Cure or Poison?'Identity Verification and the Spread of Fake News on Social Media. *Identity Verification and the Spread of Fake News on Social Media (September 14, 2018). Fox School of Business Research Paper*, (18-040).

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, 849–857.

Yuan, E. S. 2020. Zoom's Use of Facebook's SDK in iOS Client. https://blog.zoom.us/zoom-use-of-facebook-sdk-in-ios-client/, (accessed on 06/21/21).

Zhou, X.; Wu, J.; and Zafarani, R. 2020. SAFE: Similarity-Aware Multi-modal Fake News Detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 354–367. Springer.

Zhu, S.; Shi, J.; Yang, L.; Qin, B.; Zhang, Z.; Song, L.; and Wang, G. 2020. Measuring and Modeling the Label Dynamics of Online Anti-Malware Engines. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*.

Zubiaga, A.; and Ji, H. 2014. Tweet, but verify: epistemic study of information verification on twitter. *Social Network Analysis and Mining*, 4(1): 1–12.