Evaluation of Threat Information Quality Provided by Twitter

Ryu Saeki
Department of Computer Science and Engineering
Fukuoka Institute of Technology
Fukuoka, Japan
mfm22105@bene.fit.ac.jp

Kazumasa Oida
Department of Computer Science and Engineering
Fukuoka Institute of Technology
Fukuoka, Japan
oida@fit.ac.jp

***Abstract*—Twitter (currently being re-branded as "X") is widely used as a tool for collecting and disseminating information about the latest security incidents. However, the quality of threat information provided by Twitter (earliness, detailedness, and reliability) has not been studied sufficiently so far. Fresh information is required by both security experts and non-experts. Detailedness and reliability are also important criteria in two aspects. First, there are many accounts on Twitter, and anyone is free to post fake news. Second, true information that is not detailed is inconsequential to experts. This study compares the quality provided by Twitter and a security news site, which is expected to be very trustworthy. Because Emotet is having a serious impact in Japan, this study verified the earliness and detailedness by measuring when and how often words characterizing Emotet variants have appeared on Twitter and the news site in the past. Experiments revealed that Twitter alerted far earlier and more frequently about more diverse malware types, malicious attachment extensions, and malicious subject lines. Reliability was assessed based on two criteria: website reliability and text reliability. The news site was superior in terms of website reliability. In terms of text reliability, on the other hand, their difference was insignificant. The text reliability was derived from discrepancies between articles about the same security incidents, which were detected by humans and a state-of-the-art machine learning model. Overall, the quality of information on Twitter is higher than on the news site.**

Keywords*—*cyber security; Twitter; cyberthreat intelligence; threat information quality; ChatGPT.

I. Introduction

This study extends our original conference paper [1] in three aspects. First, the quality of threat information has been assessed with the addition of 2023 data. Second, web page structure containing unrelated text (e.g., advertisements, recommendations, related articles, etc.) was analyzed to collect body texts of web pages successfully. Third, an approach using Chat Generative Pre-trained Transformer (ChatGPT) [2] that automatically detects discrepancies between two descriptions was added to verify the reliability of web page content.

Twitter is a place where diverse topics are transmitted in real-time among many users. Because cybersecurity topics are also actively exchanged, many security experts are trying to extract useful Cyber Threat Intelligence (CTI) from Twitter [3]. One of the most attractive features of Twitter is its real-time nature. Because attack tactics continuously evolve and new attack techniques could suddenly emerge, cybersecurity practitioners, who want to proactively defend their organizations, are forced to retrieve timely information from Twitter.

Twitter is currently regarded as one of the major Open-Source INTelligence (OSINT) sources [4] [5]. However, except for a few studies on extracting Indicators of Compromises (IoCs) (i.e., forensic artifacts or remnants of an intrusion), quality evaluation focusing on Twitter CTI information has not been conducted [3]. Niakanlahiji et al. demonstrated the earliness of IoCs in Twitter by extracting many malicious URLs from Twitter that are not in blacklist databases [6]. Shin et al. showed the excellency in earliness, uniqueness, and accuracy of Twitter IoCs by comparing them with public CTI feeds [7].

Fake news, i.e., false or misleading information, often spreads over Twitter, especially when big events, such as the COVID-19 pandemic [8] or natural disasters [9], occur. Twitter may not always provide reliable information because anyone is free to post their own opinions. Additionally, although many security incidents are posted every day, it is not clear whether they are useful professional knowledge. The reliability and detailedness of Twitter information may not fully meet the requirements of experts and non-experts.

This study compares the earliness, detailedness, and reliability of CTI information provided by Twitter and Security NEXT [10], a major Japanese cybersecurity news site. The news site publishes daily free-access articles on security incidents and new vulnerabilities, with coverage by trusted security experts. As such, the site surely provides sufficiently reliable information with some level of detail and speed.

This study collected Japanese tweets and articles related to Emotet as a case study; Emotet is a Trojan horse that is spreading worldwide. The most common Emotet attack is to infect computer systems with various types of malware using malicious attachments in spam emails. Japan has been a major Emotet target since 2019 [11]. This study quantitatively reveals that Twitter excels in terms of detailedness and earliness but not in terms of reliability.

The remainder of this paper is organized as follows. Section II presents research related to this paper. Section III outlines security news, CTIs, and Emotet. Section IV describes our experiment and program to collect and analyzes website data. Section V presents experimental results regarding the information quality of the two media types. Section VI discusses the possibility of automatic discrepancy detection using the state-of-the-art artificial intelligence (AI) technology ChatGPT. Section VII discusses some aspects of Twitter threat information from the perspective of our findings. Finally, the conclusion of this study and future directions are presented in Section VIII.

II. Related Work

The idea of collecting CTI information from Twitter dates back more than a decade [12]. Because it is a time-consuming task due to the enormous volume of data generation, many recent studies have proposed more efficient extraction mechanisms. Some studies make use of advanced machine learning technologies [13] [5] [14] [15], some utilize semantic knowledge bases [4], and some focus on active Twitter account tracing [16] and detection [17]. Recent approaches improve the accuracy of IoCs by comparing threat reports from six sources, including Twitter [18], or by collecting tweets from non-experts who discovered or encountered attacks [19].

The term IoC is used in the computer security domain to refer to specific patterns, markers, or evidence that indicate a security incident or breach. IoCs can be represented in different forms and types depending on attack techniques. IoCs may be IP addresses, domain names, file hashes, malware behavior patterns, or network traffic characteristics. These can help detect, investigate, and respond to attacks and breaches on computer systems and networks.

In contrast, limited studies have investigated the quality of CTI information provided by Twitter. Table I compares our approach with prior studies [6] [7] [20]. All existing studies in the table extracted IoCs and compared them with blacklists or CTIs, while this study extracted the entire text that describes Emotet attacks in order to capture characteristic attack patterns and their evolution. Moreover, these studies evaluated Twitter IoCs based on earliness or timeliness. In addition to earliness, this study evaluated Twitter information based on detailedness and reliability. For comparison, we selected Security NEXT, a cybersecurity news site, which is expected to be highly reliable.

“Earliness,” “detailedness,” and “reliability” are the essential criteria that should not be missing in any one of these. It is easy to understand that earliness and reliability are indispensable; detailedness is also necessary because an easy way to increase reliability is to avoid detailed descriptions so as not to increase incorrect expressions. The earliness, uniqueness, and accuracy criteria in [7] were also treated as a set. In our opinion, Shin et al. of [7] emphasize the aspect of Twitter information diversity, while this study focuses on Twitter information volume. Reference [20] extended the work in [7] using similar quality criteria.

This study collected body text on each web page and detected discrepancies among texts using ChatGPT. The study also evaluated earliness and detailedness from three categories: malware types, attachment extensions, and email subject lines. This is because many words that appear frequently in the texts and are not always classified as IoCs fall into one of these three categories.

Table I. Comparison of existing approaches that evaluate threat information quality of Twitter.

	This study	IoCMinor [6] 2019	Twiti [7] 2021	[20] 2023
Quality criteria	earliness detailedness reliability	earliness	earliness uniqueness accuracy	timeliness overlap correctness
Extracted data	texts on Emotet	IoCs	IoCs	IoCs
Baseline	news site	blacklists	public CTIs	CTIs
Data analysis	text analysis ChatGPT	graph theory etc.	machine learning	machine learning

III. Background

A. Threat Information

Security companies and experts provide the latest threat information and advice on security measures via Twitter. They share information and engage in discussion with the hashtag #Infosec. The hashtag #Cybersecurity is also widely used and is a keyword for sharing security news, vulnerability information, and best practices.

In addition to Twitter, there are various other ways to gather threat information [18]. Security vendors and information-sharing organizations, such as Computer Emergency Response Team (CERT) and Computer Security Incident Response Team (CSIRT), provide CTI information and access to threat databases. They also provide useful early warning and countermeasure information and services to automatically collect and analyze threat information. Meanwhile, security news sites and blogs present articles and analyses on the latest threat information and attack trends regularly.

B. News vs. CTI

Both security news and CTI are important components of the cybersecurity landscape. In general, they have the following different characteristics:

- Scope:
  Security news covers a broad range of topics and provides a general awareness of current happenings in the field of security, such as the latest events, incidents, breaches, vulnerabilities, and developments. While analysis and insights may be included, they are usually limited due to the nature of news reporting. Meanwhile, CTI provides in-depth analysis, context, and actionable insights about cyber threats. It involves the collection, analysis, and dissemination of information about potential or ongoing cyber threats to help understand the tactics, techniques, and procedures (TTPs) used by attackers.
- Timeliness:
  Security news typically reports real-time or near real-time information on noteworthy security incidents, data breaches, emerging vulnerabilities, and other relevant topics. CTI focuses on long-term trends, emerging threats, and potential risks, which are not immediately visible in security news.

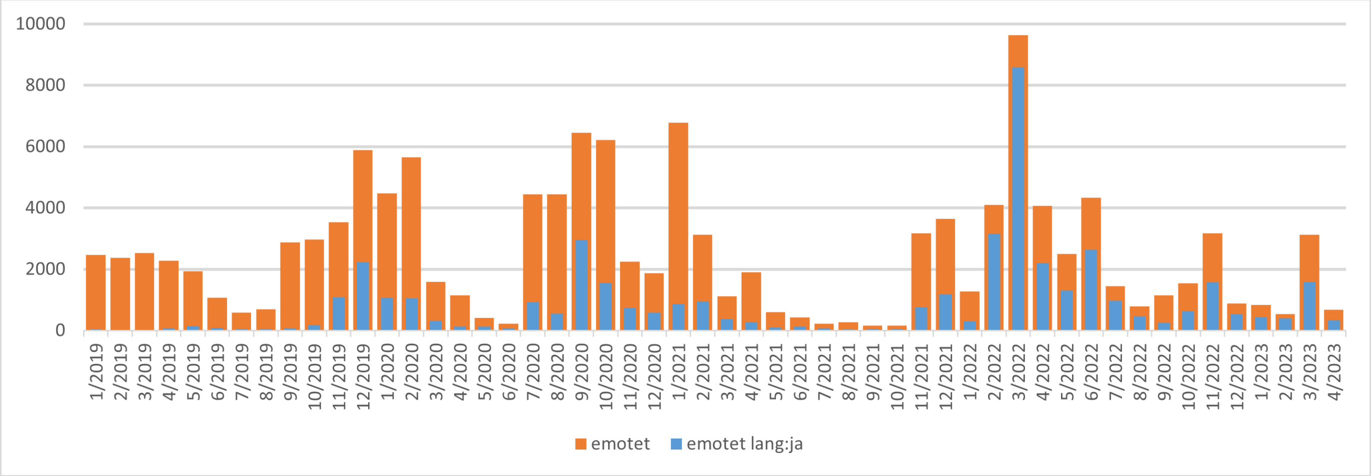


Figure 1. Number of tweets per month collected from Twitter APIs using search strings “emotet” (orange bars) and “emotet lang:ja” (blue bars). Japan is the primary target country and Emotet infection has two phases: dormant and spreading.

- Audience:
  Security news caters to a broad audience, including general users, businesses, and professionals in the cybersecurity industry. CTI is primarily targeted at cybersecurity professionals, threat intelligence analysts, incident responders, and security operations teams.

C. *Emotet*

Emotet is a type of banking Trojan first discovered in 2014. Emotet hijacks user computers through emails and opens a backdoor to download and install various malicious programs. Infected computers may be used as part of a botnet to send spam emails and spread various types of malware. The goal of the attack is to steal personal information, such as credit card information and passwords.

Figure 1 shows the monthly numbers of tweets collected with the search strings “emotet” and “emotet lang:ja,” where lang:ja limits the language to Japanese. The figure shows that Emotet has heavily affected Japan and that there were several major Emotet outbreaks, partly because of its attack strategy changes. Detailed and reliable Emotet variant information soon after its outbreak is indispensable.

Early Emotet used spam emails containing Microsoft Office files with malicious macros. Later, around 2016-17, Emotet replaced macros in spam emails with links through which users downloaded malicious files. In 2018-20, Emotet often downloaded TrickBot, a type of banking Trojan. Recently, it has been downloading ransomware Ryuk, which encrypts files on infected systems and demands a ransom.

Subjects of Emotet emails vary widely, but they generally have the following characteristics. They may disguise the content of invoices or orders, such as ”Invoice Payment Due,” contain subjects related to banks and financial institutions, such as ”Transaction Notification,” and use themes related to important documents, such as ”Urgent: Read Immediately.” Emotet can disguise file types and extensions and changes them frequently, making it difficult to identify Emotet by a specific extension alone.

IV. Software Structure

A. *Program Overview*

Figure 2 shows the structure of our Python program. Similar tools and ideas to extract IoCs from Twitter were discussed in Section II. The uniqueness of our program is to automate the process of parsing Japanese words describing Emotet threats and outputting graphs. The correctness of the output results was verified by comparing sampled outputs with manually calculated results.

As shown in Figure 2, input CSV files containing Emotet tweets are processed sequentially, characteristic Emotet words are extracted, and bar graphs are output. The following describes detailed steps 1)-6) executed by the modules in the figure (because the news site program is roughly a subset of the Twitter program in Figure 2, the two programs are described simultaneously):

1) Tweet collection: Collect all Japanese tweets containing string “emotet” using Twitter APIs. The numbers of such tweets per month are shown in Figure 1. A Google Chrome extension [21] that compiles tweets satisfying search criteria into a CSV file was used.
2) URL collection: In the case of the Twitter analysis program, collect all shortened URLs (`http://t.co/`) in the tweet, and then convert all the shortened URLs to the original URLs. Next, exclude all duplicate sites by checking whether they have the same URL, title, or text. In the case of the news site analysis program, collect URLs of all Japanese articles available on the news site.
3) Text collection: Collect text bodies (the areas enclosed by tags <p>) of web pages specified by the URLs mentioned above through scraping if the page titles include “emotet.” The reason why the areas enclosed by tags <p> are used for text body extraction is explained later in this section.

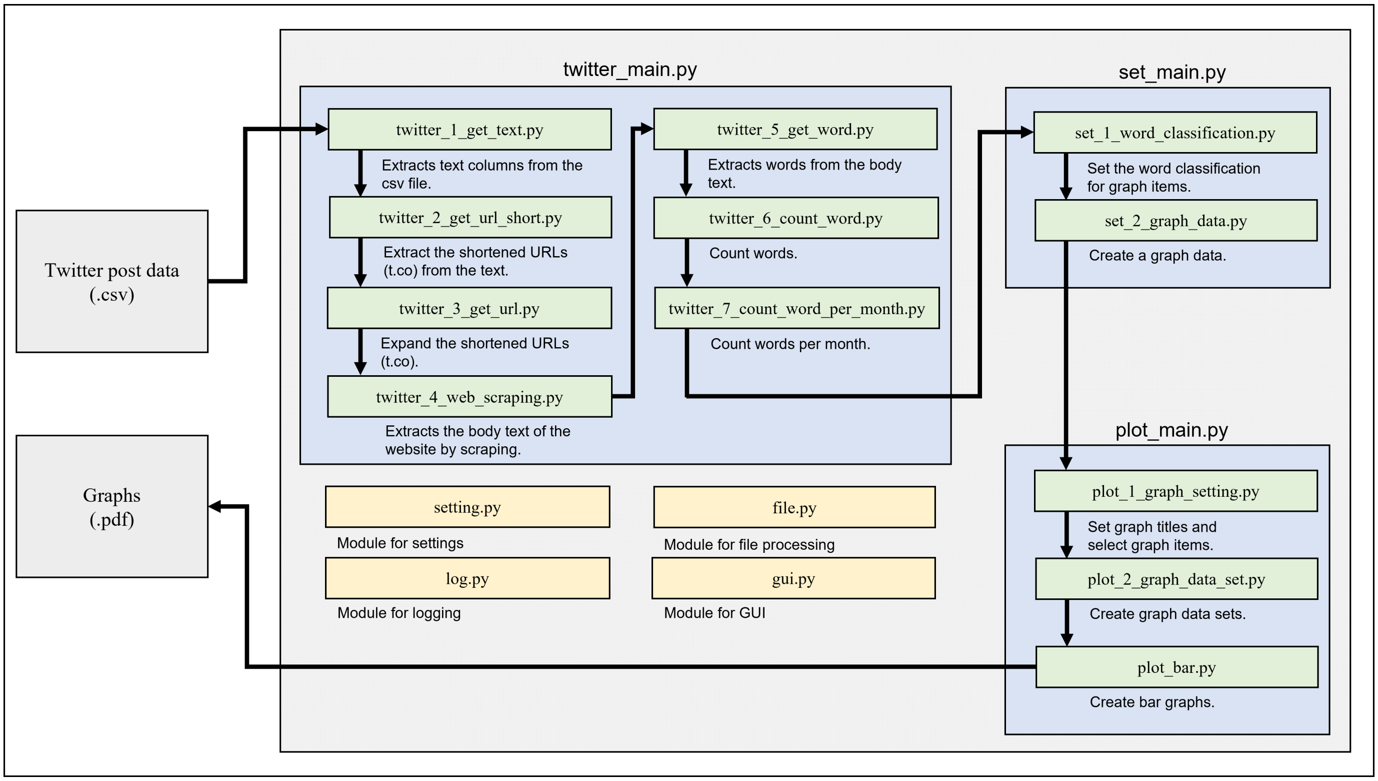

Twitter post data
(.csv)
Graphs
(.pdf)
twitter_main.py
twitter_1_get_text.py
Extracts text columns from the csv file.
twitter_2_get_url_short.py
Extract the shortened URLs (t.co) from the text.
twitter_3_get_url.py
Expand the shortened URLs (t.co).
twitter_4_web_scraping.py
Extracts the body text of the website by scraping.
twitter_5_get_word.py
Extracts words from the body text.
twitter_6_count_word.py
Count words.
twitter_7_count_word_per_month.py
Count words per month.
setting.py
Module for settings
file.py
Module for file processing
log.py
Module for logging
gui.py
Module for GUI
set_main.py
set_1_word_classification.py
Set the word classification for graph items.
set_2_graph_data.py
Create a graph data.
plot_main.py
plot_1_graph_setting.py
Set graph titles and select graph items.
plot_2_graph_data_set.py
Create graph data sets.
plot_bar.py
Create bar graphs.

Figure 2. Twitter analysis program written in Python designed to collect words about Emotet threat from Japanese tweets and output graphs.

4) Word collection: Extract nouns or compound nouns from the texts obtained above using janome [22], a morphological analysis library.
5) Word classification: Classify all collected words into several categories based on the meanings of the words. Collected words are diverse; they include a variety of synonyms and contractions.
6) Graph generation: Create graphs (e.g., histograms) that show the frequencies of characteristic Emotet words for each category.

B. Web Page Analysis

In HyperText Markup Language (HTML) [23] documents, the `<p>` tags are generally used to group text bodies into paragraphs to provide appropriate styling and semantic separators for each paragraph. Meanwhile, the `<div>` tags are used for semantic content grouping and styling and can group different types of content, so that elements other than text bodies (e.g., advertisements) can be placed within the `<div>` tags. Although the `<p>` tag is generally not used directly for advertising, it may be used, for example, to represent a caption for an image or chart or to indicate a whole or part of a quote.

Program `twitter_4_web_scraping.py` in Figure 2 extracts body texts of web pages from text areas enclosed by `<p>` tags in the HTML codes. Although `<div>` tags can also be used to enclose text bodies, they tend to be used for areas other than the body text. Table II compares the occupancy rates (how much the body is enclosed) and unrelated rates (how much the text area is not related to the body) of `<p>` and `<div>`. For comparison, the ten most frequently tweeted web pages and ten randomly selected Security NEXT news articles were used. As can be seen from the table, the mean occupancy rate is always high, independent of the tags and the media types. On the other hand, the mean unrelated rates of `<p>` for Twitter and the news site are not greater than 12.4%, whereas those of `<div>` exceed 85%. Thus, tags `<p>` were used for text body extraction. Most unrelated texts include advertisements, recommendations, related articles, etc., so statistical results obtained using `<div>` tags are strongly influenced by such texts.

C. Word Collection

Emotet spreads via spoofed emails with malicious file attachments, and installs a variety of malware on the infected devices. Accordingly, extensions of the attachments, subject lines of the spoofed emails, and types of Emotet malware are words that characterize currently popular Emotet variants and are very valuable threat information.

The program `set_1_word_classification.py` in Figure 2 collects characteristic Emotet words using regular expressions, which specify match patterns in the body texts. Table III exemplifies regular expressions for malware type TrickBot and attachment extension ONE. As shown in the table, Microsoft OneNote files correspond to those that include either one or onenote.

Table II. Number of words in the body texts, occupancy rates (percentages of the text bodies that are enclosed by <p> or <div>), and unrelated rates (percentages of text areas of <p> or <div> that are not related to the body text). The <p> tags can extract text bodies with greater accuracy than the <div> tags.

URLs in tweets	#words in body text	<p> occupancy rate	<p> unrelated rate	<div> occupancy rate	<div> unrelated rate
cybersecurity-info.com	795	100.0%	0.0%	100.0%	97.9%
cybersecurity-jp.com	6,359	95.7%	10.9%	0.0%	100.0%
htn.to	6,908	93.3%	18.1%	99.8%	78.4%
b.hatena.ne.jp	5,401	87.3%	0.5%	99.4%	57.0%
www.itmedia.co.jp	2,397	95.6%	8.0%	100.0%	93.6%
news.mynavi.jp	3,770	95.5%	10.7%	100.0%	84.2%
newsrelea.se	1,415	63.1%	3.3%	99.8%	84.0%
xtech.nikkei.com	1,755	90.0%	21.6%	99.3%	85.4%
ameblo.jp	783	37.3%	0.0%	99.5%	79.8%
hash1da1.hatenablog.com	464	4.7%	38.6%	100.0%	94.6%
Mean	3,005	76.3%	12.4%	89.8%	85.5%

Article numbers in Security NEXT	#words in body text	<p> occupancy rate	<p> unrelated rate	<div> occupancy rate	<div> unrelated rate
141138	1,068	100.0%	2.3%	100.0%	92.0%
144322	445	100.0%	6.8%	100.0%	92.7%
144577	570	100.0%	1.7%	100.0%	91.5%
136146	730	100.0%	1.8%	100.0%	100.0%
144644	311	100.0%	3.2%	100.0%	94.3%
141546	1,271	100.0%	2.0%	100.0%	90.1%
144656	354	100.0%	2.5%	100.0%	93.9%
136167	1,308	95.7%	2.1%	100.0%	90.7%
136270	1,061	100.0%	2.4%	100.0%	91.6%
144089	523	100.0%	2.1%	100.0%	91.4%
Mean	764	99.6%	2.7%	100.0%	92.8%

Table III. Regular expressions for malware type TrickBot and attachment extension ONE.

Search word	Regular expression
TrickBot	1. ^trickbot$ 2. ^trickbot[^a-zA-Z]+ 3. [^a-zA-Z]+trickbot$ 4. [^a-zA-Z]+trickbot [^a-zA-Z]+
ONE	1. ^one$ 2. ^one[^a-zA-Z]+ 3. [^a-zA-Z]+one$ 4. [^a-zA-Z]+one[^a-zA-Z]+ 5. ^onenote$ 6. ^onenote[^a-zA-Z]+ 7. [^a-zA-Z]+onenote$ 8. [^a-zA-Z]+onenote [^a-zA-Z]+

Table IV. Dataset sizes and execution times.

	Twitter	Security NEXT
Number of Websites	1,895	102
Number of words (different words)	313,814 (46,099)	6,860 (2,280)
Execution time (h)	128	8

V. Calculation Results

A. Data Size

We collected Japanese tweets and news articles describing Emotet threats posted in the period from January 1, 2019 to April 30, 2023. The numbers of Japanese tweets during this period can be seen in Figure 1. Table IV shows the number of websites we observed, the number of words (different words) in all websites, and the execution times for characteristic word extraction. The table indicates that Twitter provides larger dataset sizes; the numbers of websites and words on Twitter are 19 and 46 times greater than those of Security NEXT, respectively. The table also indicates that the number of words per website on Twitter is 2.5 times greater than that on Security NEXT. According to Table II, no news articles include more than 2,000 words, while some websites referenced from tweets exceed 6,000 words.

The execution time for Twitter was approximately 16 times longer than that for the news site. Much of the execution time was spent collecting URLs and texts (steps 2)-3) in Section IV). Therefore, the ratios of the execution times and the numbers of websites between the two media types are roughly the same.

B. Earliness

Emotet tactically distributes malware using spam emails with malicious attachments. Therefore, malware types, extensions of attachments, and subject lines of spam emails are important threat trends. Let us first verify how quickly these trends were announced via Twitter and the news site. Table V compares the dates when the two media reported each of the ten Emotet malware types [24] [25] [26] for the first time. As shown in the table, Twitter reported at least 100 days earlier for all malware types. In addition, there are four types that have not appeared on the news site yet. It is clear that security experts can more quickly share information about malware variants via Twitter.

Tables VI and VII compare the dates when the two media reported the file name extensions and email subject lines used for Emotet infection for the first time. Although Twitter provides quicker reports for these categories as well, the two categories show greater variability in delays than the malware type category. Moreover, Twitter did not issue an alert for the extension “ONE” and subject line “fire inspection” earlier than

Table V. First published dates of Emotet malware types and delays (days) of Security NEXT.

Malware	Twitter	Security NEXT	Delay
TrickBot	Apr. 13, 2019	Nov. 28, 2019	229
QakBot	Apr. 13, 2019	Oct. 8, 2020	553
Ryuk	Apr. 22, 2019	Nov. 28, 2019	220
IcedID	Apr. 13, 2019	Nov. 10, 2020	577
Zloader	Sep. 7, 2020	Dec. 22, 2020	106
Ursnif	Nov. 1, 2019	Feb. 10, 2022	101
ZeusPandaBanker	Apr. 13, 2019	unpublished	-
Gootkit	Jul. 29, 2020	unpublished	-
Conti	Mar. 22, 2021	unpublished	-
Cobalt Strike	Nov. 16, 2019	unpublished	-

Table VI. First published dates of attachment extensions used for Emotet infection and delays (days) of Security NEXT.

Extension	Twitter	Security NEXT	Delay
ZIP	May 15, 2019	Sep. 4, 2020	567
DOC	Feb. 6, 2019	Nov. 28, 2019	295
PDF	Oct. 18, 2019	Feb. 5, 2020	110
XLS	Nov. 29, 2019	Feb. 10, 2022	804
RTF	Aug. 29, 2020	unpublished	-
LNK	Jan. 23, 2020	Apr. 26, 2022	824
ONE	Mar. 16, 2023	Mar. 16, 2023	0

the news site. This result indicates that Twitter is not always the first to announce current Emotet threats.

C. Detailedness

Figure 3 illustrates the numbers of websites that described each of the ten Emotet malware types per year from 2019 to the first four months of 2023. Figs. 3 (left) and (right) correspond to Twitter and Security NEXT, respectively. Note that the vertical axis scale of Twitter is more than 10 times larger than that of the news sites. Note also that four malware types have not been reported on the news site yet. From the figure, Twitter has more websites reporting Emotet malware and more malware types than the news site for all years. Thus, Twitter provides more detailed malware information each year than the news site.

Figure 4 shows the numbers of websites that described each of the seven malicious file extensions. Again, Twitter has more detailed Emotet attachment information. For example, Figure 4 (left) indicates that Twitter has been alerting malicious XLS attachments every year since 2019, whereas from Figure 4 (right), Security NEXT reported them only in 2022.

Figure 5 shows the numbers of websites that described Emotet email subject types. Figure 5 (left) shows that at least four subject types appeared every year, while Figure 5 (right) shows that more than three types appeared only in 2020. Malware types are shared among experts, while malicious subject lines are sent to alert email users. Thus, information on Twitter is useful to non-experts as well. In summary, Twitter consistently provides far more detailed attack trends than Security NEXT in terms of malware type, extension, and subject line.

D. Reliability

This study evaluates the reliability of information from two perspectives: reliability of websites and reliability of text on the web pages. Table VIII compares eight measurements of twenty websites randomly selected from those referenced by Twitter and Security NEXT. The first six measurements are whether the website specifies the writer's name, writer's contact information, published/update date, Secure Sockets Layer (SSL) certificate [27], information source, and privacy policy. The last two verify whether the site has link errors and misspellings.

Table VII. First published dates of email subject lines used for Emotet infection and delays (days) of Security NEXT. A negative delay indicates that the news site published earlier.

Subject	Twitter	Security NEXT	Delay
Reply	Nov. 1, 2019	Nov. 28, 2019	27
COVID-19	Nov. 12, 2019	Feb. 5, 2020	85
Invoice	Feb. 6, 2019	Dec. 25, 2019	322
Bonus	Dec. 12, 2019	Dec. 25, 2019	13
Conference	Jan. 21, 2020	Jul. 31, 2020	192
Questionnaire	Dec. 20, 2019	Sep. 4, 2020	259
Fire inspection	Sep. 8, 2020	Sep. 4, 2020	-4
Christmas	Dec. 6, 2019	Dec. 25, 2020	385

Table VIII. Website reliability scores based on eight measurements. Security NEXT is distinctly superior to Twitter.

Measure item	Twitter	Security NEXT
1. Writer name	20/20	20/20
2. Writer's contact info.	17/20	20/20
3. Published/updated date	19/20	20/20
4. SSL certificate	17/20	20/20
5. Information sources	16/20	2/20
6. Privacy policy	17/20	20/20
7. No link errors	13/20	20/20
8. No misspellings	18/20	20/20
Total score	$\frac{120}{160} = 0.75$	$\frac{142}{160} = 0.89$

Table VIII concludes that Security NEXT is more reliable. The news site is perfect except that it barely specifies the source of the news articles. Although Twitter also provides a high score, reliability varies from one website to another. It should be noted that Twitter is more likely to give link errors and misspellings because, as shown in Table II, its text size is larger in most cases.

Let us next see the reliability of text on the web pages. Because it is difficult to verify whether text includes fake news, study detected discrepancies between two web pages and considered that their descriptions are not fake if there are no discrepancies. We detected discrepancies between two pages describing the same security incident, where the number of incidents we used was 53. Figure 6 shows the number of pairs of web pages describing the same security incident and the number of page pairs that include discrepancies in their descriptions. From the figure, there are eight discrepant pairs among 108 pairs referenced from different media, and there is one discrepant pair among 110 pairs referenced from Twitter. (The news site has one article per incident, so there were no discrepancy checks between news articles.) Thus, the percentage of discrepancies that occurred between the two media (between the pages referenced in tweets) is approximately 8% (1%).

Table IX shows discrepancies (highlighted in red) of the

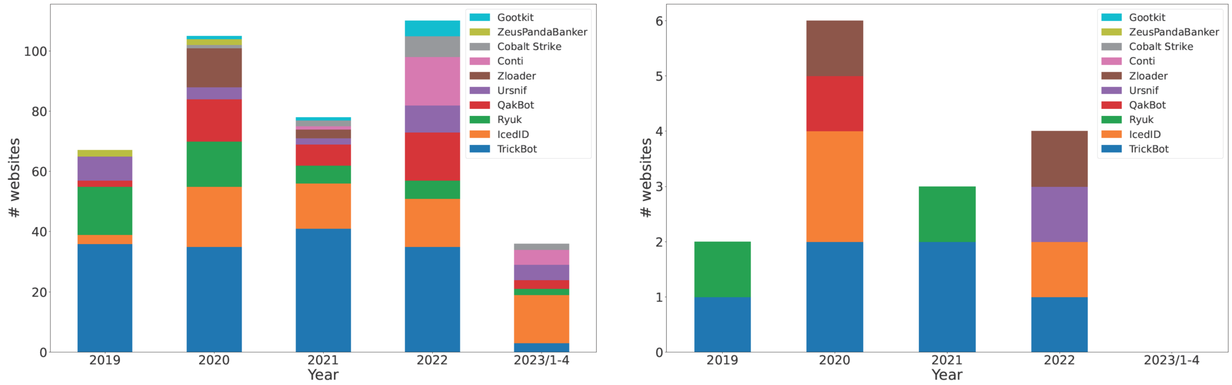


Figure 3. Yearly frequencies of Emotet malware types described on (Left) Twitter and (Right) Security NEXT.

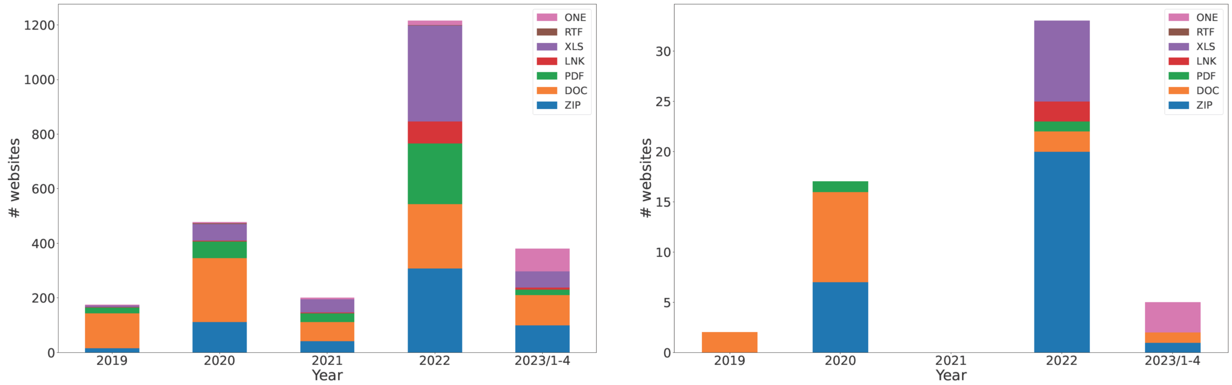


Figure 4. Yearly frequencies of Emotet attachment extensions described on (Left) Twitter and (Right) Security NEXT.

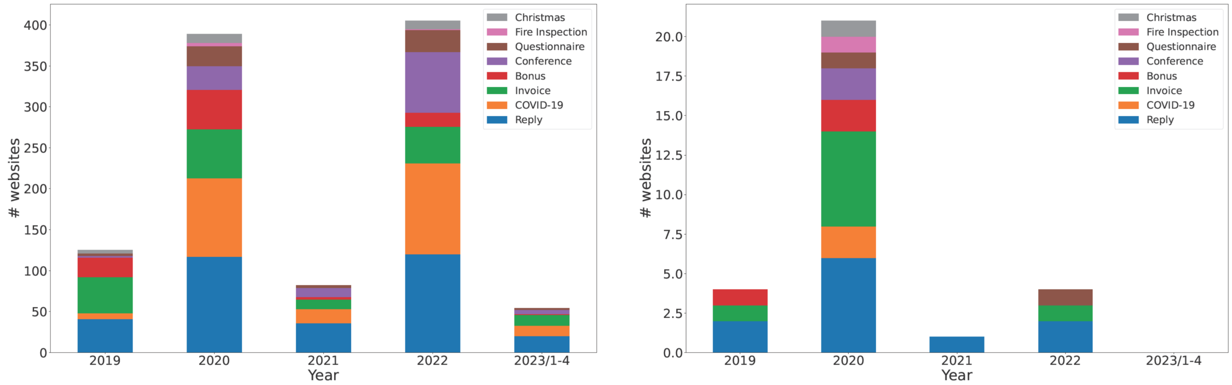


Figure 5. Yearly frequencies of Emotet subject lines described on (Left) Twitter and (Right) Security NEXT.

nine pairs. The discrepancies are all small differences in numbers and dates, except for incident 4, where two pages reported different infection routes. In other words, excluding minor numerical differences, the semantic discrepancy rate between the two media is less than 1%. Thus, the rate at which Twitter provides serious incorrect information is 1% at most.

VI. Automatic Discrepancy Detection

In the previous section, discrepancies in the descriptions of two web pages were detected by humans. When large amounts of text must be inspected, humans can make mistakes. Therefore, this section verifies whether ChatGPT can be applied to discrepancy detection. OpenAI, an American AI research laboratory, provides APIs [28] for easy access to various ChatGPT models [29] for AI developers. It is well known that slight differences in question wording could significantly alter ChatGPT's answers. Thus, we decided to develop a program that includes an OpenAI API to feed many similar questions into a ChatGPT model.

Figure 7 shows a part of our Python program. The program asks a question `question` about discrepancies between two texts: `text1` and `text2`. We prepared 144 questions with almost the same meaning. Some of them are as follows:

Table IX. Discrepancies (in red) between reports that described security incidents 1-7.

ID	Twitter	Security NEXT
1	Eighteen units were found to be infected with malware.	Eight units were found to be infected and ten were suspected.
2	On February 8, a malware infection was discovered.	On February 9, a malware infection was discovered.
3	Emails were sent to an unspecified large number of people.	Emails were sent to multiple parties.
4	The email text includes a link to an external page.	The email had an Excel file attached in macro format.
5	Emotet infection was discovered on March 15.	It was confirmed on March 16 that spoofed emails were sent after March 15.
6	2,311 leaks nationwide.	2,312 emails were leaked from infected terminals.
7	On July 11, Emotet infection was found in another terminal. On July 11, one new staff member's computer was infected.	On July 9, another employee's terminal was damaged.
ID	Twitter	Twitter
6	2,311 leaks nationwide.	A total of 2,312 emails were leaked.

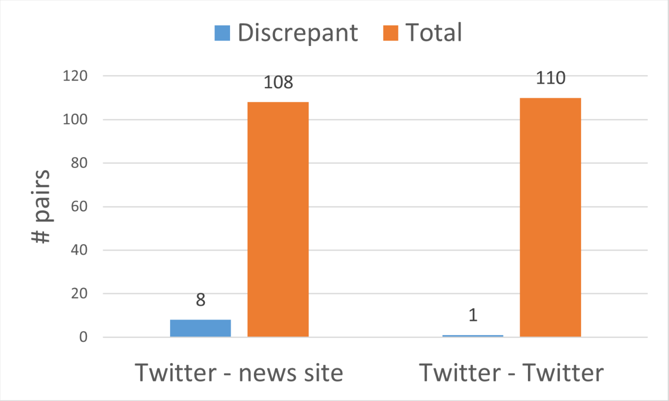
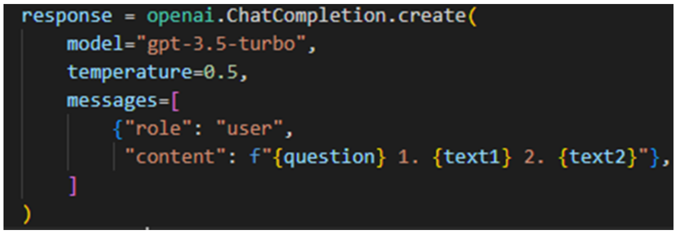


Figure 6. Among pairs of web pages describing the same incidents, there were nine pairs in which some of the descriptions do not agree. The number of security incidents was 53.



```
response = openai.ChatCompletion.create(
    model="gpt-3.5-turbo",
    temperature=0.5,
    messages=[
        {"role": "user",
         "content": f"{question} 1. {text1} 2. {text2}"},
    ]
)
```

Figure 7. Python code for receiving an answer to question `question` about `text1` and `text2` from `gpt-3.5-turbo`. A smaller `temperature` reduces output randomness (the default is 1.0).

- Are there any contradictions between the two texts?
- Please point out any discrepancies between the two texts.
- Are there any differences in the information in the two texts?
- Are there any inconsistencies in the information provided by the two texts?

Tables X and XI show the outputs of the AI program for text pairs with and without discrepancies, respectively. In the tables, TP and TN (FP and FN) denote the number of outputs detecting discrepancies correctly (incorrectly) and the number of outputs indicating no discrepancies correctly (incorrectly), respectively, and "else" denotes the number of unintelligible answers. The tables also present the percentages of these numbers among all 144 answers. Table X shows that the percentage of including correct answers (TP and TP & FP) is 17.2%. Note that the sum of TP and TP & FP for incident 2 is only 7. From this result, the 144 questions is not necessarily too many. Meanwhile, Table XI shows that the percentage of correctly answering no discrepancies is 28.7%. Thus, GPT tends to present more correct answers for text pairs without discrepancies than with discrepancies. This section has revealed that even the latest AI technology still cannot detect discrepancies automatically. In our opinion, GPT could be used as an auxiliary tool that helps reduce false positive and false negative errors.

Table X. AI program output for three text pairs (incidents 1, 2, and 7) that include discrepancies.

ID	TP	TP & FP	FP	FN	else	total
1	4	7	88	39	6	144
2	2	5	106	28	3	144
7	21	35	37	43	8	144
mean total	6.3%	10.9%	53.5%	25.5%	3.9%	

Table XI. AI program output for three text pairs (incidents 8-10) that do not include discrepancies.

ID	TN	FP	else	total
8	25	113	6	144
9	17	124	3	144
10	82	59	3	144
mean total	28.7%	68.5%	2.8%	

VII. Discussion

We have learned that it is possible to gather large amounts of up-to-date threat information about Emotet through Twitter. Furthermore, we have learned that the information is not considerably less reliable than that of the security site used for comparison. However, we also found that the following considerations must be made before collecting information via Twitter:

- Twitter alone is not enough.
  Twitter tends to provide information on Emotet malware and malicious attachment extensions much earlier than the news site. However, there was a case where the news site was quicker to warn about the subject line used in Emotet emails.
- Diverse website structures.
  Compared to the news site operated by a single organization, information on Twitter is disseminated by a diverse set of people. Therefore, when collecting information via Twitter, it is necessary to consider the structure and operating policies of the sites (e.g., update frequency,

scraping, and API). For instance, as shown in Table II, Twitter cannot always provide body text successfully using only <p> tabs, while the news site can.

- Text reliability verification.
  Table VIII suggests that Security NEXT is trustworthy, but not all Twitter users are. In other words, Twitter information should be verified in most cases. As discussed in Section VI, automatic discrepancy detection is not possible at this time, so human verification is definitely required. According to the approach in Section VI, we must consider how many questions we need to prepare depending on the GPT version and parameters. Another issue is how and how often we should find sources reporting the same security incident to perform discrepancy detection. Currently, reliability is dependent on the results of verification by some valuable sources [30].

VIII. Conclusion and Future Work

Today, many people retrieve threat information through Twitter. Because there are so many accounts on Twitter and anyone can freely transmit information, it is important to know the quality of the information provided by Twitter in advance. However, there have been insufficient studies dealing with this topic. This study measured quality based on earliness, detailedness, and reliability, which are different from the quality criteria used by previous studies.

To evaluate the quality of Twitter information, this study collected tweets about Emotet threat from January 1, 2019 to April 30, 2023, of which many tweets regarding Emotet were shared in Japan. The quality was compared with that of news articles provided by Security NEXT, a major cybersecurity news site in Japan, which is expected to be very reliable.

Our results revealed that the quality of Twitter information was far superior in terms of earliness and detailedness. However, in terms of reliability of websites, the news site achieved a better score. The discrepancy rate between descriptions of the same incidents provided by the two media was less than 1% after minor numerical differences were excluded. Accordingly, we concluded that the two media rarely contain incorrect information.

In the future, we plan to develop a fake threat news monitoring system that will regularly collect texts about the same security incidents from Twitter and news sites and publish discrepancies on an ongoing basis.

Acknowledgment

This study was supported by the Fukuoka Prefectural Police. The authors would especially like to thank Mr. Jun Koga for his helpful comments over the years.

References

[1] R. Saeki and K. Oida, "Security information quality provided by news sites and twitter," in *7th International Conference on Cyber-Technologies and Cyber-Systems (CYBER 2022)*. IARIA, 2022, pp. 42–44.

[2] "Introducing chatgpt," https://openai.com/blog/chatgpt, online; accessed 22 November 2023.

[3] M. R. Rahman, R. Mahdavi-Hezaveh, and L. Williams, "What are the attackers doing now? automating cyberthreat intelligence extraction from text on pace with the changing threat landscape: A survey," *ACM Computing Surveys*, 2021.

[4] S. Mittal, P. K. Das, V. Mulwad, A. Joshi, and T. Finin, "Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2016, pp. 860–867.

[5] L.-M. Kristiansen, V. Agarwal, K. Franke, and R. S. Shah, "Cti-twitter: Gathering cyber threat intelligence from twitter using integrated supervised and unsupervised learning," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 2299–2308.

[6] A. Niakanlahiji, L. Safarnejad, R. Harper, and B.-T. Chu, "Iocminer: Automatic extraction of indicators of compromise from twitter," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 4747–4754.

[7] H. Shin, W. Shim, S. Kim, S. Lee, Y. G. Kang, and Y. H. Hwang, "# twiti: Social listening for threat intelligence," in *Proceedings of the Web Conference 2021*, 2021, pp. 92–104.

[8] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "Fighting an infodemic: Covid-19 fake news dataset," in *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*. Springer, 2021, pp. 21–29.

[9] H. Kanoh, "Why do people believe in fake news over the internet? an understanding from the perspective of existence of the habit of eating and drinking," *Procedia Computer Science*, vol. 126, pp. 1704–1709, 2018.

[10] "Security next provides you with the latest security news daily," https://www.security-next.com/, online; accessed 22 November 2023.

[11] "The ever-changing malware "emotet" is causing more damage in japan," https://blog.trendmicro.co.jp/archives/22959, 2019, online; accessed 22 November 2023.

[12] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller, "Twitinfo: aggregating and visualizing microblogs for event exploration," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2011, pp. 227–236.

[13] U. Tekin and E. N. Yilmaz, "Obtaining cyber threat intelligence data from twitter with deep learning methods," in *2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. IEEE, 2021, pp. 82–86.

[14] N. Dionísio, F. Alves, P. M. Ferreira, and A. Bessani, "Cyberthreat detection from twitter using deep neural networks," in *2019 international joint conference on neural networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[15] V. Behzadan, C. Aguirre, A. Bose, and W. Hsu, "Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 5002–5007.

[16] A. Bose, S. G. Sundari, V. Behzadan, and W. H. Hsu, "Tracing relevant twitter accounts active in cyber threat intelligence domain by exploiting content and structure of twitter network," in *2021 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 2021, pp. 1–6.

[17] M. I. Mahaini and S. Li, "Detecting cyber security related twitter accounts and different sub-groups: a multi-classifier approach," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2021, pp. 599–606.

[18] J. Caballero, G. Gomez, S. Matic, G. Sánchez, S. Sebastián, and A. Villacañas, "The rise of goodfatr: a novel accuracy comparison methodology for indicator extraction tools," *Future Generation Computer Systems*, vol. 144, pp. 74–89, 2023.

[19] H. Nakano, D. Chiba, T. Koide, N. Fukushi, T. Yagi, T. Hariu, K. Yoshioka, and T. Matsumoto, "Canary in twitter mine: Collecting phishing reports from experts and non-experts," *arXiv preprint arXiv:2303.15847*, 2023.

[20] D. R. Arikkat, P. Vinod, R. K. A. Rafidha, A. D. Sorbo, C. A. Visaggio, and M. Conti, "Can twitter be used to acquire reliable alerts against novel cyber attacks?" *arXiv preprint arXiv:2306.16087*, 2023.

[21] "Tweet export," https://chrome.google.com/webstore/, online; accessed 22 November 2023.

[22] “Japanese morphological analysis engine written in pure python,” https://github.com/mocobeta/janome/, online; accessed 22 November 2023.

[23] D. Connolly and L. Masinter, “Rfc2854: The’text/html’media type,” 2000.

[24] “Triple threat: Emotet deploys trickbot to steal data & spread ryuk,” https://www.cybereason.co.jp/blog/cyberattack/3613/, online; accessed 22 November 2023.

[25] “Threat actor profile: Ta542, from banker to malware distribution service,” https://www.proofpoint.com/us/threat-insight/post/threat-actor-profile-ta542-banker-malware-distribution-service/, online; accessed 22 November 2023.

[26] “Threat spotlight: Panda banker trojan targets the us, canada and japan,” https://blogs.blackberry.com/en/2018/10/threat-spotlight-panda-banker-trojan-targets-the-us-canada-and-japan/, online; accessed 22 November 2023.

[27] A. Freier, P. Karlton, and P. Kocher, “Rfc 6101: The secure sockets layer (ssl) protocol version 3.0,” 2011.

[28] “Openai api,” https://openai.com/blog/openai-api, online; accessed 22 November 2023.

[29] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[30] A. Tundis, S. Ruppert, and M. Mühlhäuser, “A feature-driven method for automating the assessment of osint cyber threat sources,” *Computers & Security*, vol. 113, p. 102576, 2022.