Security Information Quality Provided by News Sites and Twitter

Ryu Saeki[1] and Kazumasa Oida[2]
Department of Computer Science and Engineering
Fukuoka Institute of Technology
Fukuoka, 811-0295 Japan
e-mails: [1]mfm22105@bene.fit.ac.jp, [2]oida@fit.ac.jp

***Abstract*—Twitter is widely used as a tool for disseminating and collecting information related to security incidents. The quality of information provided by Twitter, however, has not been studied in detail so far, where quality in this paper refers to the detailedness, real-time performance, and reliability of the information. This paper evaluates the quality of Twitter information by comparing with that of information provided by a news site that publishes a large number of security-related articles as a baseline. Our analysis showed that Twitter was significantly better in terms of detailedness and real-time performance. On the other hand, a news site was slightly better in terms of reliability.**

Keywords-Emotet; real-time performance; security; Twitter; reliability.

I. INTRODUCTION

Social Networking Services (SNS) and cybersecurity-focused news sites are two media for investigating ongoing security attacks on the Internet. SNSs can provide a large amount of fresh security information because information is transmitted in real time from a variety of sources [1]. On the other hand, news sites publish daily articles on security incidents and new vulnerabilities, with coverage by trusted security experts. In order to clarify the differences in the quality of security information provided by these two media, this study collected data on Emotet attacks in Japan provided by Twitter and Security NEXT as a case study.

Emotet is a Trojan horse that spreads primarily through spam e-mails and is raging worldwide. The most common Emotet attack method is to infect computer systems with various types of malware using malicious files attached to spam e-mails. Japan has been a major target for Emotet since 2019 [2]. Twitter includes links to external sites in its tweets and disseminates Indicators of Compromise (IoCs) from a variety of sources, including malware sandboxes, security vendor blogs, etc. It has been reported that Twitter captures ongoing malware threats, such as Emotet variants and malware distribution sites, better than other public threat intelligence [1]. Security NEXT, on the other hand, is a news site specializing in information related to security incidents in Japan, with a large number of postings and free access to all articles.

This study compared the detailedness, real-time performance, and reliability of information provided by Twitter and Security NEXT. We found that Twitter excelled in terms of detailedness and real-time performance of information about Emotet. The rest of the paper is structured as follows. In Section II, we present the process of collecting and analyzing Emotet data. In Section III, we discuss the visualized results. Finally, we conclude the work in Section IV.

II. PROGRAM STRUCTURE

We created a program that visualizes information on Emotet provided by Twitter and Security NEXT in Japanese for the period from January 1, 2019 to August 31, 2020. The following describes the program execution sequence.

1. Collect URLs posted on Twitter and Security NEXT as follows. In the case of Twitter, collect all shortened URLs (http://t.co/) in every tweet containing the term "emotet", and then convert all the shortened URLs to the original URLs. Next, check all duplicate sites (the same URL, same title, or same text) to exclude them. In the case of Security NEXT, collect all URLs contained in all of the articles of the news site.
2. Collect text areas (the areas enclosed by tag <p>) of Japanese websites of the URLs obtained above if their titles include "emotet."
3. Extract words (nouns and compound nouns) from the texts collected above using a morphological analysis library janome [3].
4. Group all collected words into several categories because they contained a variety of words including synonyms.
5. Create a histogram representing the frequency of the classified categories.

III. CALCULATION RESULTS

Table I shows the number of web sites from which information was retrieved, the total number of words, and the program execution time. Much of the program execution time is spent on the program execution sequence 1-2 in Section II. Table I shows that Twitter has more than 20 times more words than Security NEXT and that the program execution time for Twitter is more than 10 times longer. In other words, while Twitter has more detailed information than Security NEXT, it takes longer to retrieve the information.

Emotet distributes malware through spam e-mails with malicious file attachments. Therefore, we visualize the previous and current trends in Emotet's attack strategy by focusing on malware types, extensions of attachments, and subject lines of spam e-mails. Figures 1 and 2 show yearly frequencies of malware distributed by Emotet [4]-[6]. These

 ISBN: 978-1-61208-996-6

figures show that fewer types of malware appear in Security NEXT articles than in Twitter. Figures 3 and 4 show the yearly frequencies of malicious file extensions. These figures demonstrate similar result in that the ZIP format has the highest percentage in 2020, followed by the DOC format (including the DOCM format), and then the PDF format. However, in Figure 4, data for 2019 and 2021 are missing. Figures 5 and 6 show the yearly frequencies of spam e-mail subject lines. Figure 5 shows that five to six categories appear in every year, while Figure 6 shows six categories appearing only in 2020. Accordingly, Twitter tends to provide detail Emotet attack characteristics (malware, extension, subject) every year. On the contrary, Security NEXT may provide no information during the periods when Emotet attacks are less frequent.

Table II compared the dates when Twitter and Security NEXT reported the Emotet malware names for the first time. As shown in the table, Twitter reported at least 220 days earlier for all malware types. Although not mentioned in this paper, Twitter also provided quicker reports on malicious file extensions and spam e-mail subject lines.

Table III compares Twitter and Security NEXT on eight reliability measures of 20 randomly sampled websites. Here we measure the reliability of information on Twitter and Security NEXT based on website reliability. The table shows that Security NEXT is slightly better. Twitter is dependable in that it mostly has links to information sources, but the probability of link errors is not negligible. Twitter sometimes does not include writers' contact information and privacy policy statements.

TABLE I. DATA SET SIZES AND EXECUTION TIMES.

	Twitter	Security NEXT
Number of Websites	1,660	91
Number of different words	42,347	2,091
Execution time (h)	63	6

TABLE II. DATES MALWARE NAMES WERE FIRST REPORTED.

Malware name	Twitter	Security NEXT
TrickBot	Apr. 13, 2019	Nov. 28, 2019
QakBot	Apr. 13, 2019	Oct. 8, 2020
Ryuk	Apr. 22, 2019	Nov. 28, 2019
IcedID	Apr. 1, 2019	Nov. 10, 2020

TABLE III. COMPARISON FROM EIGHT RELIABILITY MEASURES.

	Reliability measure	Twitter	Security NEXT
1	Writer name	20/20	20/20
2	Writer's contact info.	13/20	20/20
3	Published/updated date	20/20	20/20
4	SSL certificate	20/20	20/20
5	Information sources	14/20	1/20
6	No link errors	8/20	20/20
7	No misspellings	18/20	20/20
8	Privacy policy	13/20	20/20
Total		113/160	141/160

IV. CONCLUSIONS AND FUTURE WORK

Today, security experts significantly depend on Twitter information. This paper quantitatively evaluated the quality of Twitter information in terms of detailedness, real-time performance, and reliability. Our results showed that the quality of Twitter information was excellent in terms of detailedness and real-time performance. On the other hand, a news site was slightly better when measured based on reliability. In the future, we will evaluate the reliability of Twitter information using other methods such as language-based and knowledge-based approaches.

REFERENCES

[1] 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.

[2] "The ever-changing malware "EMOTET" is causing more damage in Japan," 2019. [online]. Available from: https://blog.trendmicro.co.jp/archives/22959. [retrieved: August, 2022].

[3] Japanese morphological analysis engine written in pure python". [online]. Available from: https://github.com/mocobeta/janome/. [retrieved: August, 2022].

[4] "Three threats: TrickBot deployment by Emotet and data theft and Ryuk proliferation by TrickBot," 2019. [online]. Available from: https://www.cybereason.co.jp/blog/cyberattack/3613/. [retrieved: August, 2022].

[5] "Threat Actor Profile: TA542, From Banker to Malware Distribution Service," 2019. [online]. Available from: https://www.proofpoint.com/us/threat-insight/post/threat-actor-profile-ta542-banker-malware-distribution-service. [retrieved: August, 2022].

[6] "Threat Spotlight: Panda Banker Trojan Targets the US, Canada and Japan," 2018. [online]. Available from: https://blogs.blackberry.com/en/2018/10/threat-spotlight-panda-banker-trojan-targets-the-us-canada-and-japan. [retrieved: August, 2022]

 ISBN: 978-1-61208-996-6

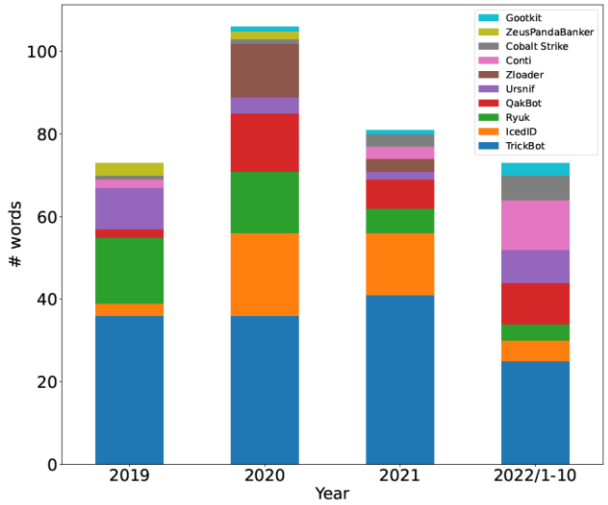

Figure 1. Malware distributed by Emotet (Twitter)

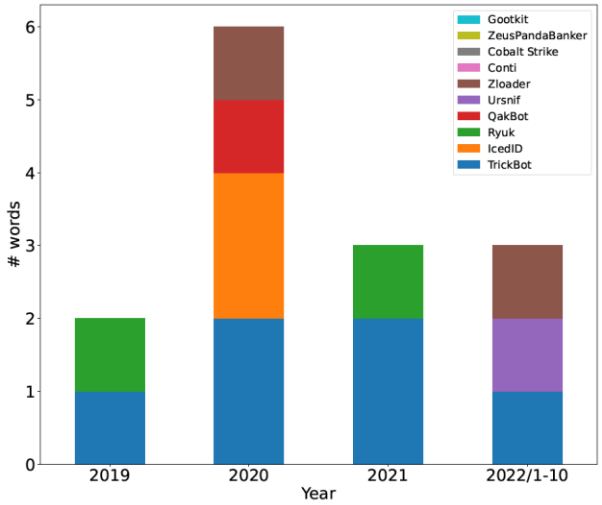

Figure 2. Malware distributed by Emotet (Security NEXT)

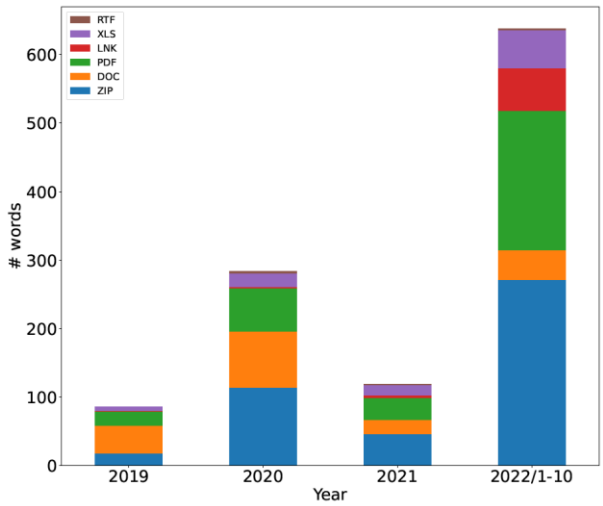

Figure 3. Malicious file extension (Twitter)

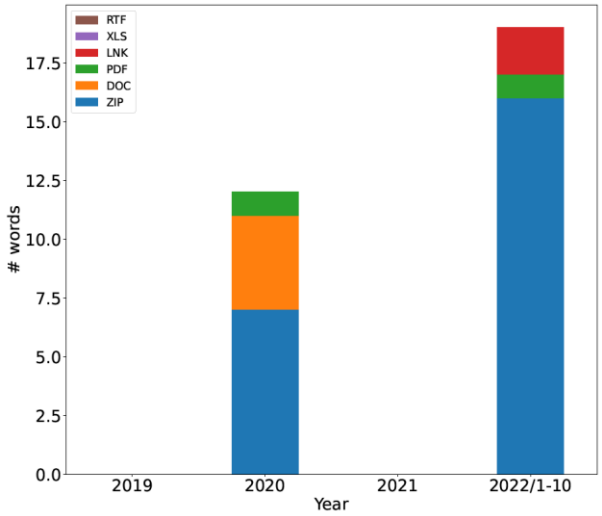

Figure 4. Malicious file extension (Security NEXT)

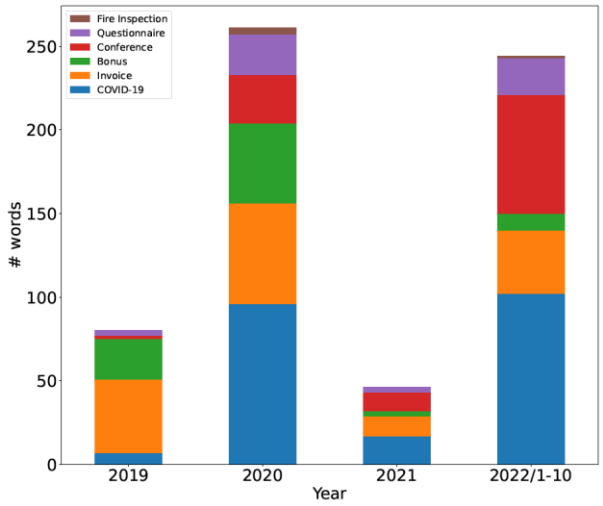

Figure 5. Spam e-mail subject line (Twitter)

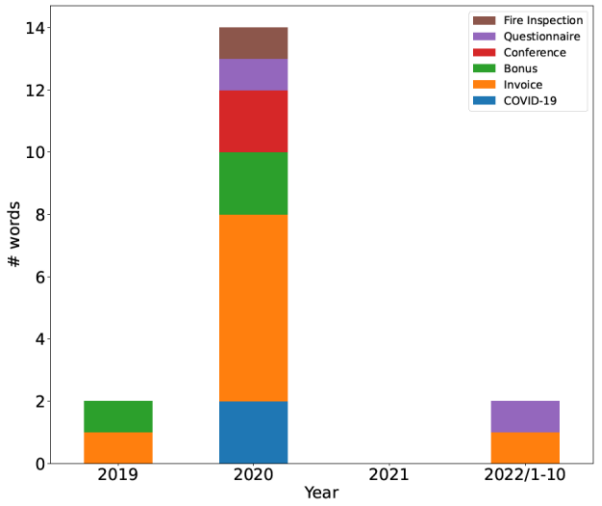

Figure 6. Spam e-mail subject line (Security NEXT)

 ISBN: 978-1-61208-996-6