IoCDigger: Extracting Indicators of Compromise (IoCs) from Social Media

Reginald Harper
rharpe10@uncc.edu
University of North Carolina
Charlotte, North Carolina

Amirreza Niakanlahiji
aniakanl@uncc.edu
University of North Carolina
Charlotte, North Carolina

Bei-Tseng Chu
billchu@uncc.edu
University of North Carolina
Charlotte, North Carolina

ABSTRACT

Each year, companies and organizations incur billions of dollars in loss due to cyber attacks. To lessen the impact of these attacks, companies have started to share their cyber threat information with the public in hope to increase awareness about ongoing attacks and also to get the benefit of the data shared by other providers. Twitter, a microblogging social media platform, is one of the most commonly used platform for sharing such cyber threat information. In this poster, we propose a novel pipeline, called IoCDigger, to efficiently extract basic indicators of compromise (IoCs) such as malicious IP addresses and URLs from tweet stream. IoCDigger utilizes various text mining techniques to locate and extract IoCs from the tweet stream. We demonstrate the effectiveness of IoCDigger by extracting IoC from a tweet dataset; it achieves 97.2% accuracy on this dataset.

KEYWORDS

Indicator of Compromise, Cyber threat intelligence, Data extraction, machine classification

ACM Reference Format:
Reginald Harper, Amirreza Niakanlahiji, and Bei-Tseng Chu. 2019. IoCDigger: Extracting Indicators of Compromise (IoCs) from Social Media . In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In cyberspace, attackers have the opportunity to study their targets and craft attack plans accordingly, while defenders must fortify their systems against all possible attacks without truly knowing about the attackers or their plans. This information asymmetry between attackers and defenders makes the defending task significantly more resource demanding and highly susceptible to failure. Sharing information about recent cybersecurity incidents can considerably reduce this knowledge gap as many techniques and also resources such as network resources are reused in different attacks to reduce the operational costs. In recent years, many cybersecurity companies and professionals have started to share cyber threat intelligence, in particular indicators of compromise (IoCs), about ongoing attacks to the public in hope to increase awareness about these attacks and also to take benefit from the shared information about those attacks that were out of their sight.

Twitter, a microblogging platform, is one of the most commonly used social media by cybersecurity professionals to share such threat intelligence with the public. However, finding useful information on Twitter is a challenging problem as it has more than 500 million active users who tweets about 100 million times per day; making this problem tantamount to looking for a needle in a haystack.

In this work, we propose IoCDigger, a novel pipeline to efficiently locate and extract IoCs from the Twitter platform. IoCDigger utilizes various text mining techniques to identify tweets containing IoCs and also to extract useful information from them. We demonstrate the effectiveness of IoCDigger by extracting IoCs from Twitter over a span of one week. Based on our experimentation, it can achieve a high level of accuracy, 97.2%.

2 PROCESS AND METHODOLOGY

In IoCDigger, as it is shown in Figure 1, we perform three major steps, namely data collection, filtering and normalization, and machine classification.

2.1 Data Collection

Only a tiny percentage of tweets shared on Twitter contains IoCs; thus analyzing all of them to find IoC is not efficient. To improve efficiency, IoCDigger collects tweets posted by known security professionals that regularly post IoCs. To be specific, IoCDigger collects tweets posted by a predefined list of Twitter users.

2.2 Normalization and Vectorization

IoCDigger performs several filtration and normalization steps on the collected tweets. It, first, tokenize the collected tweets, turning them into bags of words. Next, each bag is filtered by removing stop words and then stemming the words. Finally, the remaining words were formatted into a frequency distribution. The words in the distribution were ultimately used as features to help classification.

2.3 Classiftcation and Extraction

The third and final step in this pipeline is performing machine classification on the vectorized tweets to identify the one that contains IoCs. In addition, such tweets are matched against a set of regular expression rules to extract basic IoCs such as IP address and URLs. IoCDigger utilizes RandomForest algorithm to recognize tweets containing IoCs. To be more specific, it constructs a classifier by training the Random Forest classifier on a set of already-labeled

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HOTSOS19, April 02–03, 2019, Nashville, TN
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9999-9/18/06. . . $15.00
https://doi.org/10.1145/1122445.1122456

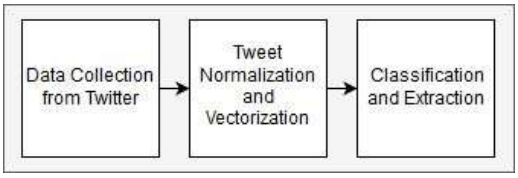


Figure 1: IoCDigger Overall Architecture

tweets. The features in our work are the words that constitute the tweet contents, number of hashtags, and number of mentions in the training dataset.

3 RESULT

4 DATASET

To evaluate our proposed pipeline, we monitored 75 accounts that are known to share cybersecurity for a week in April 2018 and collected their tweets. We gathered 2300 tweets in this way. It is noteworthy that not all of such tweets are security related as they may also tweet about other matters such as their personal life. We manually labeled each of the collected tweets as having IoCs or not having IoCs. During this process, we removed the ones that were not entirely in English.

4.1 Preliminary Results

To evaluate IoCDiger, we first trained the classifier with a set of 200 random tweets and tested a group of 143 random tweets, all from the same week. The machine classification of this group yielded an accuracy of 97.2%. In the second experiment, we again trained the classifier on 200 tweets, but this time, we made sure that 20% of the tweets were labeled as containing an IoC. When testing the 143 tweets once more, the accuracy was, again, around 97%.

We also examined the extracted artifacts, supposedly IoCs, to see whether they are actually IoCs. To this end, several weeks after data collection, we checked the extracted IoCs against VirusTotal, a web application that scans suspicious files and URLs to detect different types of malware. We found that of the fourteen IoCs checked, eleven of them were indeed IoCs, meaning that most of the tweets were correct in their identification as threats. This is important, as it shows that many IoCs could be reported on Twitter before they are reported on VirusTotal, giving Twitter an advantage over such anti-virus websites.

4.2 Misclassiftcation

Despite achieving a high level of accuracy, 97.2%, we found the results partly concerning as the misclassified tweets were all false negatives. Originally being labeled by the human as having IoCs, these tweets were estimated incorrectly by the machine as not having IoCs. The reason why these tweets were misclassified is yet to be determined, but it will be an issue that would be examined further as we continue.

5 RELATED WORK

While several research works have been proposed to extract useful cyber threat intelligence from unstructured text such as APT reports, only a few works have been proposed to extract such information from social media, in particular, Twitter.

In this section, we review a closely-related research work proposed and compare it with our project. In [1], Sabottke et al. examined data published on Twitter to identify tweets that discussed zero-day vulnerabilities; the ones that had never been publicly disclosed by the vendors. In their study, they only considered tweets that contained a specific set of hashtags, such as #vulnerability. As a result of this study, they identified a group of 4,335 users who posted information-rich messages about real-world exploits.

While the social media platform examined by both their project and ours is the same, the approaches taken were vastly different. Sabottke et al. mainly tracked hashtags, along with the keyword ”cve” For our project, we did not track any keyword, rather, we took every word passed in and based classification on the words passed in. Further, the specific issues at the heart of each project were different. In Sabottke et al’s project, the issue to be addressed was that of CVE reporting and discussion. This explores a different scope within the larger landscape of vulnerability reporting, as in our case, we searched for IoC reporting. While both projects aim was to explore vulnerability and exploit reporting on Twitter, the approaches and details of each were different.

6 FUTURE WORK

In order to make this tool most efficient, there would be more steps to take. One of the biggest steps that would need to be taken is that of consistency over time. This accuracy is great; however, it only pertains to one week’s worth of tweets. To track the consistency of the tool over time, we would conduct a longitudinal study that would compare the accuracy with the data we have used against the accuracy with data over longer periods of time, such as a week, month, or year from the original training set. This would allow us to know the stability of the classifier and would reveal how often the training data needs to be updated.

7 CONCLUSION

In conclusion, we presented IoCDigger, a pipeline to extract indicators of compromise from social media, in particular, Twitter. IoCs could be missed by VirusTotal or MetaScan at one period, but picked up a few weeks later, yet all the while was on Twitter. By discovering this treasure of IoCs being reported through Twitter and building a pipeline that can accurately detect tweets containing them, we have built a measure by which the same organizations and everyday people that are threatened can protect themselves by detecting the threats earlier. Through experimentation, we demonstrate that this pipeline can achieve a high accuracy. This pipeline could be a great contribution to the cybersecurity community and could change how IoC reporting is done.

Acknowledgement, this work is supported in part by NSF grant 1266360

REFERENCES

[1] Sabottke, C., Suciu, O., and DumitraÈŹ, T. Vulnerability disclosure in the age of social media: exploiting twitter for predicting real-world exploits. In *24th {USENIX} Security Symposium ({USENIX} Security 15)* (2015), pp. 1041–1056.