Electronic Preprint for Journal of Information Processing Vol.28

Regular Paper

Social Media Data Mining for Proactive Cyber Defense

Ariel Rodriguez[1,a)] Koji Okamura[1,b)]

Received: June 15, 2019, Accepted: November 29, 2019

Abstract: The Internet is constantly evolving, producing many new data sources that can be used to help us gain insights into the cyber threat landscape and in turn, allow us to better prepare for cyberattacks. With this in mind, we present an end-to-end real-time cyber situational awareness system which aims to retrieve security-relevant information from the social networking site Twitter.com. This system classifies and aggregates the data extracted and provides real-time cyber situational awareness information based on sentiment analysis and data analytics techniques. This research will assist security analysts in rapidly and efficiently evaluating the level of cyber risk in their organization and allow them to proactively take actions to plan and prepare for potential attacks before they happen.

Keywords: data mining, social networks, security, security and society, text mining, web intelligence, security, sentiment analysis, machine learning, machine learning & data mining

1. Introduction

The Internet has made the world increasingly more connected [1] bringing with it an increase in the number of cyberattacks [2]. Organizations often deal poorly with this issue, employing outdated defensive strategies [3], reacting to attacks rather than proactively finding ways to prepare for threats before they happen [4]. Defensive strategies are not ideal since they place the targets of the attack a step behind malicious actors [5].

Luckily the internet is constantly evolving, and many new open and closed data sources have become available. Data resources that provide open source intelligence such as social networking sites, security news sites and blogs can be utilized to help plan and prepare for threats. From the various social networking sites, Twitter is a particularly interesting and useful platform for security companies, security researchers, hacktivist groups, and other related entities.

These entities often use social media and Twitter to post news on new vulnerabilities and types of attacks. Hacktivist groups and hackers have also used Twitter as an organizational method to inform followers of rallies, coordinate cyberattacks and as a fear mongering system to disseminate threats [6], [7]. This shows that Twitter is a "treasure trove" of information that provides unique social context compared to other data sources such as network traffic or flags and alerts.

An example of the value in mining Twitter for security information can be seen in the case of a malicious actor who uses the alias SanboxEscaper. This developer released a zero-day vulnerability which they disclosed on Twitter as well as linking to proof-of-concept exploit code on Github. Within two days of disclosing the vulnerability, a threat group known as PowerPool gang modified the proof-of-concept code for their own malicious purposes and began a hacking campaign using the vulnerability [8], [9]. If we can efficiently retrieve and process such relevant security tweets, we can then analyze large amounts of data to gain a better understanding of possible threats.

To do this we develop a unique real-time system which utilizes data analytics on Twitter data to aggregate large amounts of tweets and generate cyber situational awareness information from that data. This system also takes advantage of the context contained in tweets by performing sentiment analysis and uses that data to gain insights into threat risk level. Finally, we aggregate all our data points and provide a visualization in a form which is easy to understand and can be rapidly acted upon. This system can be used to; easily consume large amounts of relevant data points, rapidly triage potential threats, and proactively plan or prepare for cyberattacks before they happen.

Paper Organization Section 2 describes related research which has some connection or relation with the components of our study. Section 3 looks at the design and methods used for the components which make up this system. Section 4 will look at the results of each of the components and Section 5 will discuss the results of these components. Section 6 will go through a case study showing a real-world example of how this system can be used and the usefulness that can be gained from it. Finally, in Section 7 we go over the conclusions of this paper.

2. Related Work

In this section, we look at research that is specifically related to our work.

The usefulness and quality of this system's output is based on its input data. There are various studies that have used tweets as a data source for security analysis [10], [11], [12] and even more that have used tweets as a general social predictor [13], [14], [15]. In these studies the tweets that are processed are mined based on

[1] Kyushu University, Fukuoka 819–0395, Japan
[a)] roda@kyudai.jp
[b)] oka@ec.kyushu-u.ac.jp

tweet hashtags. Hashtags are a type of keyword that users can add to their posts that can help classify them and make them viewable to users who are interested in those keywords. Our research aims to gain more granularity by scanning the body of the text as well as hashtags for specific security words which are relevant to cybersecurity. Our research also adds upon the previous method by adding a second level of filtering to remove false-positives due to terms with double meanings. Using these methods allows us to obtain a larger number of better quality posts allowing our output to also be of higher quality.

In Hernandez-Suarez et al. [16] they aim to predict cyberattacks using a social sentiment sensor in Twitter. That is done by scraping tweets and putting them through a machine learning model which classifies their sentiment as positive, negative or security. The number of security or positive/negative tweets are then compared to historical data of large events during the 2016 US presidential campaign. This study measured the total number of tweets and used them as an indicator to detect possible cyberattacks. This research has shown that we can gain important insights into the cybersecurity domain from Twitter sentiment analysis by analyzing historical data. In our case, we apply tweet mining methodologies to obtain tweets and additionally utilize that data to create a system that takes advantage of the real-time nature of Twitter. Our system also allows for more fine-grained analysis of events by allowing analysts to obtain specific details about tweets and thus obtain more precise insights about those events.

In Mittal et al. [17] they present CyberTwitter, CyberTwitter is a framework that analyzes tweets and outputs alerts that can be used by analysts or other systems. This system uses a named entity recognizer trained on CVE descriptions to identify relevant entities in tweets and then uses various ontologies to map strings to real-world instances. If intelligence that may be of interest to the user is found an alert is issued with the specific details. In Lee et al. [18] they developed a web service called Secbuzzer for exploring emerging cybersecurity topics by mining open threat intelligence such as Twitter and RSS feeds. Secbuzzer creates a security expert network by ranking users that are deemed highly involved in the cybersecurity domain and then leverages this community for its content. Secbuzzer then uses an emerging topic finding algorithm to demonstrate an outline of security threats close to what happens in a realistic environment.

In both cybertwitter [17] and Secbuzzer [18] they use data mining techniques on open source intelligence with the aim of anticipating or generating relevant threat intelligence. We also mine open source intelligence although in our case we use a multi-stage filtering method which passes through tweets in multiple stages to classify them rather than a named entity recognizer. This method is explained in detail in Section 3.1. This research also uses sentiment analysis as a metric in our system to add a new dimension that can be analyzed and used as an indicator to identify or prioritize threats.

Furthermore, our research considers how security analysts utilize tools. Fink et al. [19] conducts a study of cybersecurity professionals and visualizations in a work environment. They look into the types of tools that security analysts use and how they use those tools in their daily routines. Based on this research they then reveal design principles to make effective visualizations with security analysts in mind. We have applied some of these principles such as; making the main interface work well with large displays, allowing for linkages between data abstractions and raw data and not greatly obfuscating our data. This was done to ensure that our tool not only produces useful outputs but is also a useable and effective workspace for cyber analysts.

3. System Design & Methodology

3.1 Data Mining and Filtering

Our first step in our system is to develop a method to retrieve legitimate data which is related to the cybersecurity field. Retrieving and processing non relevant data in our system can negatively affect our results and is not ideal. The data obtained using this method will then be stored to create a cybersecurity tweet specific dataset which can be used for analysis and for training our classification model.

We first filter tweets based on user accounts. In Hernandez-Suarez et al. [16] well identified Twitter accounts related to Hacktivists, security researchers and enthusiasts are identified. For our account filtering, we expand upon this list by adding more recent and pertinent accounts. By using these accounts we can confidently assume the data retrieved will be relevant and legitimate.

We subsequently filter retrieved tweets based on a security keyword filter. Even though data is filtered by security-related accounts, these accounts are not guaranteed to post purely security-related content. In Lee et al. [17] clustering analysis is used on security accounts showing that these accounts also post about sports, movies, and other non-related events. Because of this, we further filter the tweets by relevant security terms. If the tweet contains a term from our security words list such as “ddos” or “mitm” it is considered to be relevant data. In Al-Rowaily et al. [20] the “2011 Analyst’s desktop binder” is used which has a section outlining keywords to be used when monitoring social media to “provide situational awareness”. We use this as a basis for our security words list, but also expand upon it with new attacks and attack types.

When utilizing this technique it is still possible to retrieve unrelated data. Some security terms have dual meanings or different meanings in another language. For example, the word backdoor can be used in two contexts, meaning we have to distinguish between the two. We do this by analyzing the body of text against our “security-related words list”. This is a list of the most frequently used terms in our corpus which are not already in the

Table 1 Top 10 terms in our created dataset by word frequency.

Word	Frequency
scammers	13,181
virus	12,900
worm	12,468
hacker	10,522
exploit	8962
malware	7393
phishing	6755
new	6623
cybersecurity	5443
infosec	4372

Algorithm 1: Pseudocode showing the algorithm used to filter normal tweets to security-related tweets.

```
Input  : Tweets from "security user accounts" list
Output: English tweets related to cybersecurity
if tweet contains word in security words list then
    if word is in "double meaning words list" then
        if tweet contains word in "security related words list" then
            store tweet
        else
            drop tweet
        end
    else
        store tweet
    end
else
    drop tweet
end
```

security words list. This was obtained by processing our dataset using Term Frequency and analyzing the results. By doing this we obtained a list of words that are commonly associated with threats such as "injection" "attack" and "hit". In Algorithm 1 we show the flow of how we utilize this multi-stage filtering method to obtain purely security-related tweets.

3.2 Updating Dataset

Once we have our dataset, we add an updating component allowing it to integrate new security terms based on Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF). This allows for new attacks to be incorporated into our word filters which in turn allows the system to remain relevant and useful over time.

New attack names and types appear regularly in the cybersecurity domain, for instance, the term "wannacry" referring to the ransomware attack was not considered security related until after the attack was released. Therefore, it could not have been on any watchlists or filters prior to its naming and hence required a mechanism to become aware of its new context.

By retrieving tweets using our filtering system there is a high likelihood that new attack names will appear in our dataset by association. By identifying these new attack names and terms we can add them to our security word filter to improve the quality of data we retrieve. To achieve this we process our corpus using TF and TF-IDF and combine the resulting lists, excluding stop words and joining words. If one of the terms in the resulting list is found to be relevant it can then be incorporated into the existing security words list, and data containing this word will be incorporated into our dataset. Term selection is currently done manually since the TF/IDF process does not ensure that all terms will be relevant and adding non relevant terms to the filter can affect the accuracy of the system. Regularly performing this process on our corpus allows us to have an updating dataset that can stay up to date with new attack terms. A depiction of this process is shown in **Fig. 1**.

3.3 Sentiment Classifier

The third component is to create a subsystem that accurately classifies the tweets we scrape with a sentiment of negative or positive. Having accurate information about the sentiment of the

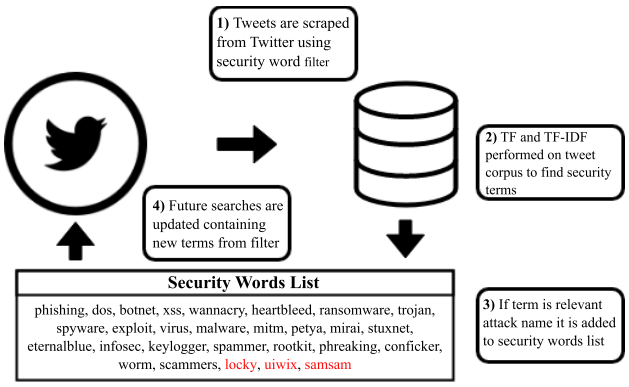

Fig. 1 The flow of the updating dataset method.

tweets we scrape allows us to make better-informed decisions about our data. Experiments using domain specific lexicons show that they can improve sentiment classification, especially in domains like cybersecurity [21]. We utilize our own security dataset as a training corpus to achieve improved sentiment classification.

The sentiment140 [22] dataset is a general tweet dataset which we use for non relevant data. This dataset was annotated using a lexicon method where tweets are classified with a score based on the amount of positive or negative words contained in the tweet. For our dataset we add sentiment annotations using the same method implemented through the python library textblob. The sentiment scores are normalized for both datasets to 1 or 0 for security related and non security related data respectively.

Having this information gives us a new dimension that we can consider when looking into our data. For example, a large number of negative tweets that contain "DDoS", can be utilized as a indicator for an analyst to investigate into a "DDoS" attack occurring. This is reinforced by Ref. [16], who find a correlation between negative social media sentiment and cyberattacks. It is possible for positive tweets to contain information about a vulnerability but positive tweets that contain security terms are more likely to be used in a promotional sense such as advertising an article, product or service which isn't as pertinent to an analyst. In **Table 2** we show an example of this with positive and negative security tweets.

We use a machine learning method to do sentiment analysis since it allows us to utilize our security dataset. We trained various machine learning classifiers such as Naive Bayes Classifier, Support Vector Machine and Decision Tree using both security and non security datasets to judge the results.

3.4 Situational Awareness Interface

Finally, we visualize the various components to create an interface that shows the data obtained. This system is able to bring together large amounts of data points often referred to as "big data" [23] and present it in a way that is easy to understand and can be quickly acted upon.

4. Results & Implementation

4.1 Data Mining and Filtering

We scanned 1,716,787 total tweets during our testing phase for this research process. After removing duplicate entries and

Table 2 Positive and negative tweet examples from our system containing security terms.

Positive Tweets
Good Guys Win: Cops Hijack Botnet, Remotely Wipe Malware From 850,000 Computers https://t.co/ePl...
If Looking For A Great VPN That Is DDOS Proof And Cheap Per Month HMU. If Need A Connect For A Botnet And Puller Also HMU. Cheap Services
Negative Tweets
BLOCK .@<user>..he sends DM's with links to malware. numerous times! I hate people like him.....
Microsoft: We're fighting Windows malware spread via Excel in email with bad macro https://t.co/UJY... via @<user>; @<user>

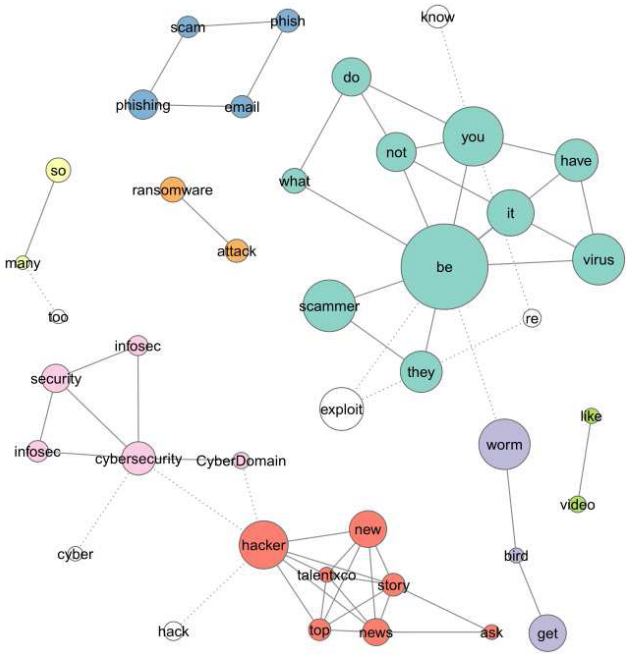

Fig. 2 Co-occurrence networks for the created dataset.

scanning for security words the final dataset contained 70,475 domain-specific entries. **Table 1** shows the top 10 words in the dataset by frequency, we can see that 9 of the top 10 words are security related. This helps to show that the dataset has a large number of entries from our security word list and that our filtering method was able to obtain relevant data.

To further show that our dataset is domain specific, we can look at **Fig. 2**. This figure depicts the co-occurrence network [24] of terms in our dataset. Looking at this figure we can see graphs which show groups of co-occurring terms. Each relevant subgraph contains a security word from our list as a main hub of the graph. All the relevant relationships and interconnections of our dataset revolve around attack types which shows that the dataset is security related and therefore specific to our domain.

4.2 Updating Dataset

Performing TF and TF-IDF on our corpus produced two lists of relevant terms. From these lists, we found three entries of interest. These terms were Locky, Uiwix and TSMC. Locky is the name of a ransomware type delivered as a Microsoft word attachment, Uiwix is a type of ransomware that uses Eternalblue to infect systems, and TSMC is the acronym for Taiwan Semiconductor Manufacturing Company which is a semiconductor company that was hit by a Uiwix attack. These terms fit our criteria of being attack names which were not currently in our filter, this shows our method was able to assist in the process of identification.

Table 3 Score comparison for classification models trained and tested using the general tweet dataset.

Classifier	Precision	Recall	F1 Score
Logistic Regression	0.80	0.80	0.80
Linear SVM	0.79	0.79	0.79
Multinomial NB	0.77	0.77	0.77
Bernoulli NB	0.78	0.78	0.78
SGD Classifier	0.78	0.78	0.78
Decision Tree	0.72	0.72	0.72

Table 4 Comparison of evaluation metrics for logistic regression classifier trained and tested using various datasets.

Dataset	Precision	Recall	F1-Score
Created Dataset	0.85	0.84	0.84
General Dataset	0.82	0.82	0.82
Half Created Half General Dataset	0.83	0.83	0.83

Table 5 Comparison of evaluation metrics across datasets with balanced polarities of varying sizes.

	Precision	Recall	F1-Score
79,000	0.79	0.79	0.79
200,000	0.81	0.81	0.81
400,000	0.81	0.81	0.81
800,00	0.82	0.82	0.82
1,200,000	0.82	0.82	0.82

4.3 Sentiment Classification

Table 3 shows the results of training various classifiers using the general tweet dataset with 80% of the dataset used for training and 20% for testing. The logistic regression model performed the best. Logistic regression is a method of fitting a regression curve to variables and then making a prediction based on those values, in our case we use bag-of-ngram features as those values. These features are extracted from the entire tweet corpus as an input to our logistic regression model. Further tuning of the ngram parameter improved results with ngram range 1, 2 giving an f1 score of 80% and the shortest amount of training time.

Table 4 shows the results of training our logistic regression model on three datasets, once again using 80% of the dataset for training and 20% for testing. We use our created dataset containing security data, the general tweet dataset which does not contain security data and a balanced dataset containing both. Our dataset containing security data achieves the best results, by training our classifier on data that is similar to what will be used in the system we are able to more accurately classify our data, and achieve higher quality output.

In **Table 5** we show the classification results of the general dataset for varying dataset sizes where each dataset is balanced

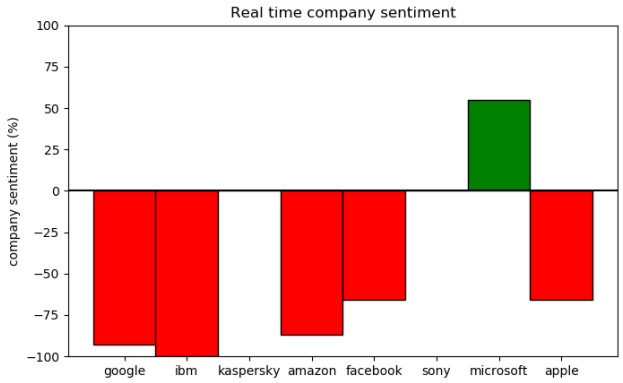

Fig. 3 Interactive user interface for system showing average sentiment scores of tracked companies. Clicking on sentiment bar opens company tweet detail window.

google	-93.3333
google total tweets	15
google neg tweets	14
google pos tweets	1
google total tweets w/sec words	2
google neg tweets w/sec words	2
google pos tweets w/sec words	0
google sec words #	phishing [4, 4]

Fig. 4 Company tweet detail window displayed when entity in main interface is selected showing statistics about tweets captured for selected company.

based on polarity, this means we have the same amount of negative and positive entries in the dataset. In Table 5 we can see that as the size of the dataset used for training increases so do the classification scores.

4.4 Situational Awareness Interface

In **Fig. 3**, we show the main interface window of this system. It displays the aggregated general sentiment of all scanned tweets in a clear and easily readable way. Each column presents the total average sentiment output for a specific company shown by its label. In Ref. [25] they outline some design principles for security visualizations such as avoiding 3D graphics, avoiding complex visualizations that require explanations and providing aggregation of data that is easily readable. We have applied those principles and used a simple layout and common color scheme with positive sentiment being in green and negative being in red. This is a common layout theme and color scheme allowing any analyst utilizing this to quickly get an idea of the real-time sentiment towards an entity at any given time. This type of interface theme also lends itself to use on large displays which is useful in operation centers and was a design recommendation in Ref. [19].

If the user wishes to investigate further into a particular entity based off of the output of the interface, they can select the appropriate column to display further details. **Figure 4** shows the detail window which is presented to the user when a column is selected from the general sentiment window. This pane presents the details such as the total amount of tweets, the sentiment values for those tweets and the security words contained within those tweets.

5. Evaluation & Discussion

5.1 Data Mining and Filtering

We filtered our data based on security accounts to ensure a level of legitimacy, by doing this we are limiting the effect of fake and erroneous data on our system. We then analyzed our dataset and found that it contained relevant cybersecurity related tweets showing that our multi-stage filtering method successfully achieves its goal. This system's output is dependent on the data that is inputted into it, fake, erroneous or non-relevant data can skew visualizations to present incidents which are not correct, therefore the incorporation of these methods is important to the quality and credibility of the system.

5.2 Updating Dataset

The relevance of having an updating dataset for our system is important for various reasons. Firstly, due to the fast-paced nature of the security industry new attack names and types appear frequently, to be able to defend against these attacks it is important to first be aware of their existence. By updating our keyword filter regularly and accurately we can ensure these attacks are incorporated into our system and reflected in its output.

5.3 Sentiment Analysis

In this component of the system, we create a sentiment classification model using machine learning to classify our tweets with a value of positive or negative. We chose to use a machine learning method due to its ability to automatically classify data based on a labeled dataset which allows us to make use of our own created dataset. When classifying data based on niche subjects it is beneficial to use domain-specific training data due to the idiosyncrasies of that area. In Table 4 we show that having a full domain specific dataset for training can provide better results over more general datasets.

The usefulness of a machine learning model is often tied to the amount and variability of data that is contained in its dataset. Training our model on the greatest amount of examples possible will improve its classification since it is exposed to more diverse data, this is shown in Table 5 where the general tweet dataset obtains better results as the size of the dataset it is trained on increases. This reinforces a well-known notion in machine learning that more data often produces better results. At this point our created dataset has 79,064 entries and is able to achieve 85% accuracy as shown in Table 4. This shows that one area we can achieve an accuracy gain is by increasing the size of our dataset.

Lastly, the amount of positive or negative entries contained in the dataset can have an effect on the classification accuracy of the model. Having too much or too little of either can mean misclassification. Because of this, we aim to have a similar amount of positive and negative entries as would appear while scanning real-time Twitter data. Our dataset has 60% positive entries and 40% negative entries since this dataset was retrieved from real-time Twitter data it shows an accurate break up of positive and negative occurrences. Because of this, we have implemented a balanced dataset which ranges from 50% to 60% positive to achieve best results.

5.4 Situational Awareness Interface

This section shows the main interfaces of our system and explains how they can be used to interpret the data of our system. These interfaces have been created with the intent of providing a clear representation of the data points we acquire and allow the user to acquire more detailed data when it is required. Section 6 will look at specific real-world examples of how this system and its interfaces can be utilized to provide value to security analysts, groups or users.

6. Case Study

Here we show a real-world example of how our system can be used to assist in rapidly and efficiently analyzing large amounts of user-generated data.

To make the output of our system more useful and relevant we filter tweets based on a specific organization. In this example, we have filtered out all tweets that pass through our system that relate to the company Google. Google maintains a huge amount of IT infrastructure and online services, due to this giant footprint the amount of vectors that malicious users can use to attack Google is huge. This system can assist in protecting Googles assets, users and employees through greater threat awareness.

Figure 5 shows the median general sentiment value recorded from this interface between the dates of 2019/3/26–2019/4/20. It shows that Google received negative general sentiment for the whole testing period. The lowest general sentiment value is −55% on 2019/4/18 and the highest is −88% on 2019/3/29, meaning the values are within a 33 point range. There are areas in the graph where multiple days build up, creating a peak, such as from 2019/3/26 until the peak of 2019/3/29 and also from 2019/4/3 until 2019/4/5.

In Ref. [19], data visualizations are used to assist security analysts in understanding data. Their analysis shows that analysts "would typically investigate spikes to determine what caused the feature". Therefore these "spikes" indicate relevant points where analysts would investigate further. When looking further into the details of a specific event a useful metric is negative tweets that contain security words.

Figure 6 shows a comparison of general sentiment score to negative tweets containing security words. In this graph, we can see that similar peaks are visible and more pronounced compared to the general sentiment score. That is to say that naturally as the number of negative tweets rise, the general sentiment score falls. Similarly to the general sentiment graph, we have visible peaks where there is a greater amount of negative tweets containing security words. An analyst can use this information as an indicator to look into the days or times which have a higher percentage of negative tweets containing security words.

Fink et al. [19] also explains an analyst need to be "able to drill down and get as much details as possible when needed". Therefore the natural progression for an analyst would be to look into which words were most prevalent in tweets during these peaks. **Figure 7** shows the occurrences of specific security words compared with negative tweets containing those words. In this figure the peaks are clearly defined and are created from the amount of negative tweets rising at those specific points.

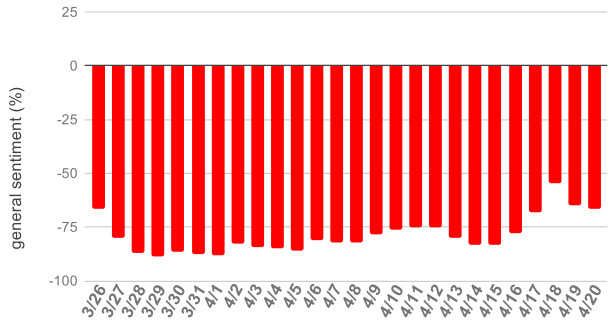

Fig. 5 General sentiment values by day over the testing period for Google.

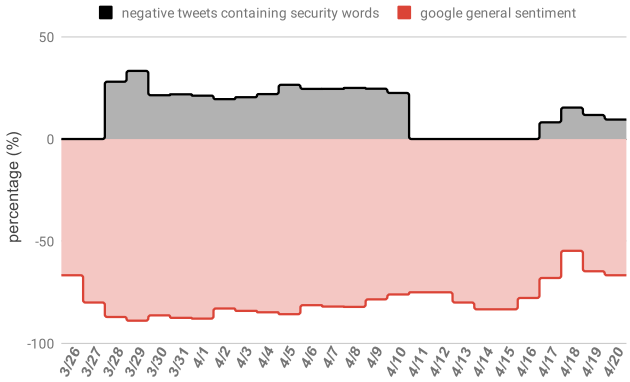

Fig. 6 General Sentiment (red) vs negative tweets containing security words (black).

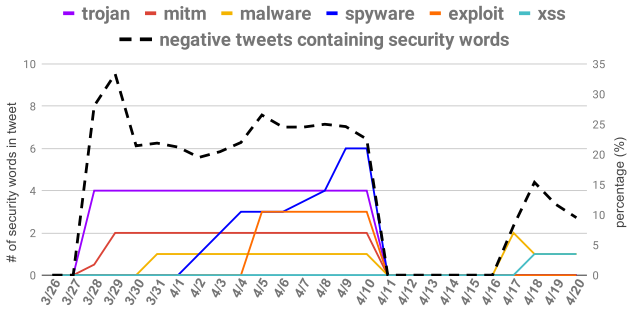

Fig. 7 Occurrences of security words in negative tweets vs percentage of negative tweets of total tweets.

Finally, the analyst can drill down to the original tweets themselves to gain further context into the threat as is depicted in **Fig. 8**.

In the case of the first peak, the main driver was the word trojan. This was due to various tweets about a popular web browser called "UC Browser" hosted on the Google Play store which has an ability to download new libraries and modules to the user's phone without permission [26]. The occurrence of mitm further drove this trend to its peak. The tweets containing mitm were also referencing "UC Browser" and its ability to be used to perform mitm attacks and download new plugins.

In the second peak, we saw that the rise correlates with the word spyware and then is exacerbated by the rise of the word exploit. After checking the specific tweets containing the word spyware we saw that they are related to a spyware application called "Exodus" that was once again hosted on the Google Play Store. Exodus is a type of spyware which imitates a legitimate application and exfiltrates data from various applications and services on Android phones. The spyware is alleged to be developed

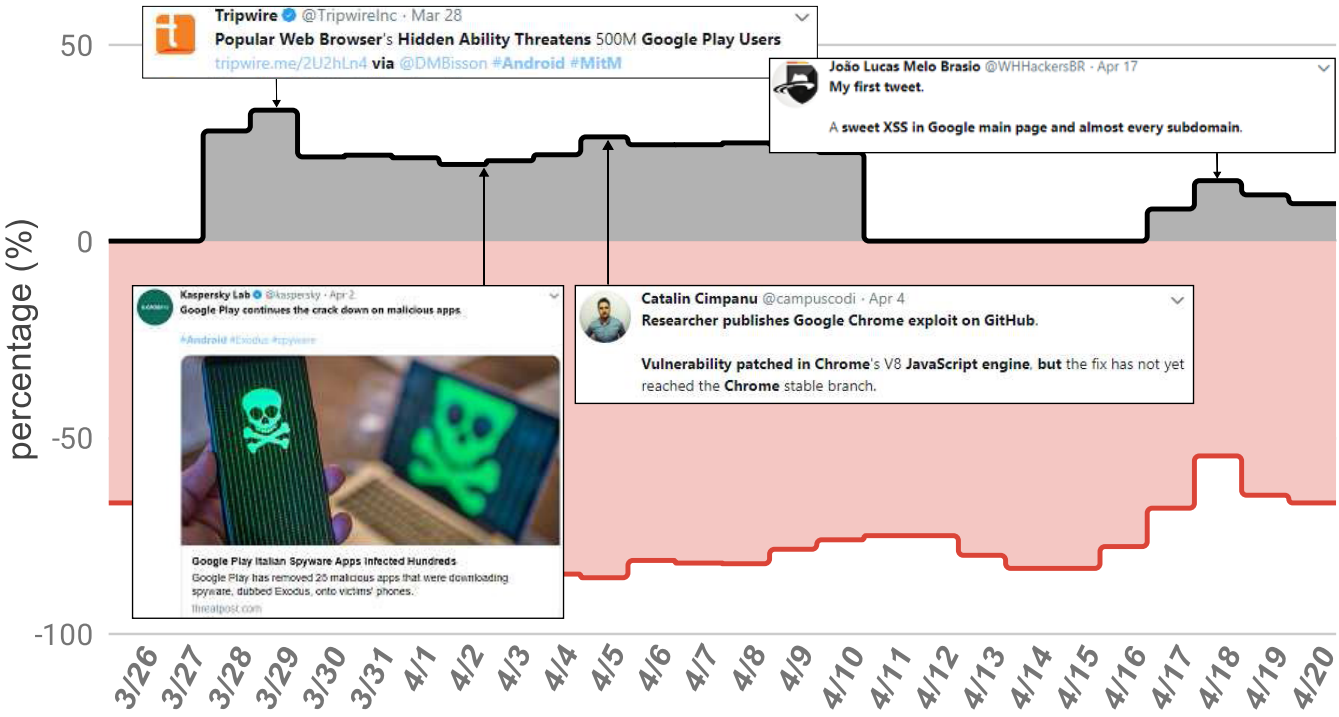

Fig. 8 Example of particular tweets which can be searched during peaks to gain detailed insight into specific threats.

by the Italian firm eSurv and has links to the Italian Government, with its goal being to gather information on Italian citizens [27].

The second peak rose further with the presence of the word exploit. After analyzing the relevant tweets this came with the presence of tweets referencing a security researcher named István Kurucsai who published a vulnerability along with its proof of concept for an un-patched Google Chrome Vulnerability.

Finally the third peak which rose on the appearance of the term xss came from a tweet posted by an individual security researcher which had found an xss vulnerability in “Google main page and almost every subdomain”.

From these examples it is easy to see how analysts can benefit from the enhanced awareness that this system is able to provide. In the following section, we will discuss how this system compares with other methods and how it can help various types of users prepare for attacks.

6.1 Comparison & Mitigation

Our system improves and expands upon conventional and traditional methods used by security analysts. One of the most, traditional and widely used methods is described in Ref. [19], “One tactical analyst described his daily routine by examining a collection of blogs and websites to find out the new vulnerabilities.” “From these reports he identifies a list of approximately 50 threats that he needs to examine that day. He prioritizes this list to determine the top 10 that must be addressed. Of these he is usually able to tackle the top 3–5 during the day.”

Identifying, examining and prioritizing threats manually is very time consuming and inefficient. In an industry where speed is crucial, our system is able to take advantage of the benefits of social media to improve upon previous conventions. To show this we compare our results to the popularity of Google searches

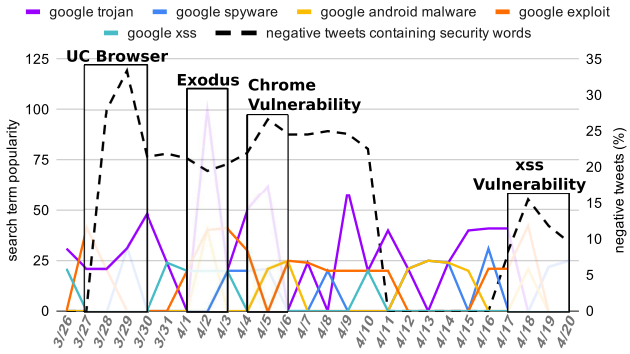

Fig. 9 Our negative tweets measurement shows our method has greater sensitivity and faster detection compared to Google search popularity.

performed for the terms we track. It is common for analysts to investigate threats using online search engines and Google searches make up 50% of external traffic to online publishers [28]. Therefore by looking at Google search popularity we can compare the time difference between when our system detects and prioritizes a threat and the traditional method of reading articles online and investigating threats manually using search engines. We show the results in **Fig. 9** where we highlight search popularity for each term during the same periods.

Our system encountered the term trojan in conjunction with Google for the first time on 2019/3/27, these terms were used in tweets referring to “UC Browser” a malicious application hosted on Google’s Play store platform. We can see that search popularity for the term google trojan rose on the 30th, this shows that awareness of this threat became widespread in the community on the 30th which is three days after our systems first detection. Our system prioritized the importance of this threat based on a rising negative tweet measurement on the 28th, two days before search popularity hit its peak in this time window. Also, an entry for the

"UC Browser" vulnerability was published in the NIST national vulnerability database (CVE-2019-10251) on 2019/3/28, one day after we first detected the term trojan in our system.

Looking at tweets related to the "Exodus" malicious application, our system detected the term spyware on 2019/4/2. Google search popularity for the term google spyware peaks on the 3rd, one day after our systems first detection. Our negative tweet measurement also rises during this period indicating increasing importance of this threat.

Our negative tweet measurement continued to rise due to tweets containing Google and exploit in relation to a Google chrome vulnerability. Our first detection for these terms came on the 4th while Search popularity for google exploit falls to zero on the 5th, and rises the next day showing a postponed public reaction. The fix for this chrome exploit was released on 2019/4/30, 25 days after our first detection.

We attributed the third peak to a single tweet from a security researcher regarding a Google chrome xss vulnerability. In this case there was no increase in search popularity, a tweet from a single user without a large number of followers can easily be lost within the greater news cycle. These types of tweets hold a great amount of value since they are published straight to social media rather than security news sites.

Finally, we looked at 6 popular cybersecurity news sites to compare against our system. In 12.5% of the cases the cybersecurity sites released an article about the same threat we had tracked before our system was able to detect it. In 33.3% of cases our system detected the threat before one of the sites published an article regarding the same threat. In 54% of cases the sites did not publish an article regarding the threat detected by our system.

Our system detected threats before they were published on certain security websites or became known in the wider community as well as detected threats which did not receive any attention on news sites. Our system prioritized these threats using our negative tweet measurement allowing analysts to focus on mitigating threats before any damage is caused. Our system can provide speed, efficiency and effectiveness benefits over traditional and conventional methods.

7. Conclusion

In this paper, we have presented a real-time cyber situational awareness system that provides cybersecurity-related information that can be useful to security analysts and users for preparing and planning for cyberattacks. We developed a filtering method that retrieves cybersecurity relevant data from security accounts on the website Twitter.com. We were able to make use of this data to train a machine learning model and show the benefits of integrating security data into our dataset. Finally, we created an interface that aggregates large amounts of data and applies design principles specific to cybersecurity workspace visualizations so that security analysts can easily and efficiently utilize our system to evaluate the current cyber risk level in their organization and proactively defend themselves from threats.

For future work, this system can incorporate other social network services, as well as forums, and other open source intelligence sources. By increasing the scope of sources into the system we can track more data points and be aware of emerging trends in other platforms. We will also work on classifying the data that is taken into our system. Within our data there are various subgroups, identifying these subgroups and labeling them will help us to better find the value of each tweet.

References

[1] Manyika, J.: The internet of things: Mapping the value beyond the hype (2015).

[2] Abomhara, M. et al.: Cyber security and the internet of things: Vulnerabilities, threats, intruders and attacks, *Journal of Cyber Security and Mobility*, Vol.4, No.1, pp.65–88 (2015).

[3] Rowe, B.R. and Gallaher, M.P.: Private sector cyber security investment strategies: An empirical analysis, *The 5th Workshop on Economics of Information Security* (*WEIS06*) (2006).

[4] Miller, K.L.: What we talk about when we talk about "reasonable cybersecurity": A Proactive and Adaptive Approach, *FLA. BJ*, Vol.90, pp.23–23 (2016).

[5] Lewis, J.A.: Raising the bar for cybersecurity (2013).

[6] Uitermark, J.: Complex contention: Analyzing power dynamics within anonymous, *Social Movement Studies*, Vol.16, No.4, pp.403–417 (2017).

[7] Coleman, E.G.: Logics and legacy of anonymous, *Second International Handbook of Internet Research*, pp.1–22 (2018).

[8] Seals, T.: Active spy campaign exploits unpatched windows zero-day (2018).

[9] Seals, T.: Windows zero-day drops on twitter, developer promises 4 more (2019).

[10] Burnap, P., Javed, A., Rana, O.F. and Awan, M.S.: Real-time classification of malicious URLs on Twitter using machine activity data, *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (*ASONAM*), pp.970–977, IEEE (2015).

[11] Gupta, B., Sharma, S. and Chennamaneni, A.: Twitter sentiment analysis: An examination of cybersecurity attitudes and behavior (2016).

[12] Mohammad, S.M., Kiritchenko, S. and Zhu, X.: NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets, arXiv preprint arXiv:1308.6242 (2013).

[13] Conover, M.D., Gonçalves, B., Ratkiewicz, J., Flammini, A. and Menczer, F.: Predicting the political alignment of twitter users, *2011 IEEE 3rd International Conference on Privacy, Security, Risk and Trust and 2011 IEEE 3rd International Conference on Social Computing*, pp.192–199, IEEE (2011).

[14] Sinha, S., Dyer, C., Gimpel, K. and Smith, N.A.: Predicting the nfl using twitter, arXiv preprint arXiv:1310.6998 (2013).

[15] Takahashi, B., Tandoc, Jr., E.C. and Carmichael, C.: Communicating on twitter during a disaster: An analysis of tweets during typhoon haiyan in the philippines, *Computers in Human Behavior*, Vol.50, pp.392–398 (2015).

[16] Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Perez-Meana, H., Olivares-Mercado, J. and Sanchez, V.: Social sentiment sensor in twitter for predicting cyber-attacks using l1 regularization, *Sensors*, Vol.18, No.5, p.1380 (2018).

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

[18] Lee, K.-C., Hsieh, C.-H., Wei, L.-J., Mao, C.-H., Dai, J.-H. and Kuang, Y.-T.: Sec-buzzer: Cyber security emerging topic mining with open threat intelligence retrieval and timeline event annotation, *Soft Computing*, Vol.21, No.11, pp.2883–2896 (2017).

[19] Fink, G.A., North, C.L., Endert, A. and Rose, S.: Visualizing cyber security: Usable workspaces, *2009 6th International Workshop on Visualization for Cyber Security*, pp.45–56, IEEE (2009).

[20] Al-Rowaily, K., Abulaish, M., Haldar, N.A.-H. and Al-Rubaian, M.: Bisal–a bilingual sentiment analysis lexicon to analyze dark web forums for cyber security, *Digital Investigation*, Vol.14, pp.53–62 (2015).

[21] Thelwall, M. and Buckley, K.: Topic-based sentiment analysis for the social web: The role of mood and issue-related words, *Journal of the American Society for Information Science and Technology*, Vol.64, No.8, pp.1608–1617 (2013).

[22] Go, A., Bhayani, R. and Huang, L.: Twitter sentiment classification using distant supervision, *CS224N Project Report, Stanford*, Vol.1, No.12, p.2009 (2009).

[23] Ward, J.S. and Barker, A.: Undefined by data: A survey of big data definitions, arXiv preprint arXiv:1309.5821 (2013).

[24] Higuchi, K.: KH Coder 3 reference manual, Kioto (Japan): Ritsumeikan University (2016).
[25] McKenna, S., Staheli, D. and Meyer, M.: Unlocking user-centered design methods for building cyber security visualizations, *2015 IEEE Symposium on Visualization for Cyber Security* (*VizSec*), pp.1–8, IEEE (2015).
[26] BISSON, D.: Popular web browser's hidden ability threatens 500m google play users (2019).
[27] Seals, T.: Exodus: New android spyware made in italy (2019).
[28] Trielli, D. and Diakopoulos, N.: Search as news curator: The role of google in shaping attention to news information, *Proc. 2019 CHI Conference on Human Factors in Computing Systems*, p.453, ACM (2019).

Ariel Rodriguez was born in 1986. He received his B. Eng. (Hons) degree from RMIT University in 2012. He joined Kyushu University in 2015 and received his M.IS degree in 2019. He is currently a doctoral student at Kyushu University. His research interests are the application of machine learning to solve real world cybersecurity problems.

Koji Okamura received his B.S., M.S. and Ph.D. from Kyushu University in 1988, 1990 and 1998, respectively. He became an associate professor of the Computer Center and Graduate School of Information Science and Electrical Engineering in 1998 and a professor at Kyushu University in 2011. He serves as the director of the Cybersecurity Center at Kyushu University and vice director of the Research Institute for Information Technology, and vice CISO of Kyushu University. He is a member of IPSJ, IEICE, IEEE-CS.