On the Automated Assessment of Open-Source Cyber Threat Intelligence Sources

Andrea Tundis[1], Samuel Ruppert[2], and Max Mühlhäuser[1]

Department of Computer Science, Technische Universität Darmstadt (TUDA),
Hochschulstrasse 10, 64289, Darmstadt, Germany,
{tundis,max}@tk.tu-darmstadt.de
Deutsche Bahn AG, Frankfurt am Main, Germany
samuel.ruppert@deutschebahn.com

Abstract. Global malware campaigns and large-scale data breaches show how everyday life can be impacted when the defensive measures fail to protect computer systems from cyber threats. Understanding the threat landscape and the adversaries' attack tactics to perform it represent key factors for enabling an efficient defense against threats over the time. Of particular importance is the acquisition of timely and accurate information from threats intelligence sources available on the web which can provide additional intelligence on emerging threats even before they can be observed as actual attacks. In this paper, an approach to automate the assessment of cyber threat intelligence sources and predict a relevance score for each source is proposed. Specifically, a model based on meta-data and word embedding is defined and experimented by training regression models to predict the relevance score of sources on Twitter. The results evaluation show that the assigned score allows to reduce the waiting time for intelligence verification, on the basis of its relevance, thus improving the time advantage of early threat detection.

Keywords: Open source cyber threat intelligence · Cybersecurity · Machine Learning · Feature engineering · Twitter.

1 Introduction

Emerging vulnerabilities in computer systems can lead to far reaching impacts due to the high number of possibly affected systems. Cyber Threat Intelligence (CTI) is an emerging field whose main mission is to research and analyze trends and technical developments related to Cybercrime, Hactivism and Cyberespionage, based on the collection of intelligence using open source intelligence (OSINT), social media intelligence, human intelligence. Current research directions are exploring OSINT as a means to proactively gather CTI from individuals and organizations that share relevant information (e.g. vulnerabilities, zero-day exploits) publicly on the web, sometime just spread, sometime to openly recruit groups (so called "hacktivists") for an imminent attack campaign [13]. This scenario, and the fact that timeliness is essential in security, emphasizes the

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_34

need to determine the relevance of such information not only based on whether it is already widely spread but also on the quality and informativeness of the source itself [17]. Different publishers, security professionals, vendors and researchers provide cyber threat-related information on vulnerabilities and even hackers post information about ongoing attack campaigns or new vulnerabilities on social media like Twitter, as well as forums and marketplaces in the darkweb. Obviously, this information varies strongly with regards to credibility, timeliness and level of detail, and it is difficult to acquire and assess it in an automated manner since the sources do not only vary content-wise but also regarding their structure and syntax. To understand these evolving threats, it is essential for security experts to illuminate the threat landscape including adversaries, their tools and techniques [9]. To deal with this need, it is simply not practical to implement counter-measures in a timely and economical manner for all possible attacks, but learning about the details of cyber threats relevant sources and, prioritizing them is a vital step in defending computer systems.

For this reason the extraction of CTI from such open sources, i.e. publicly accessible data on the internet, has been the target of recent research in the field of OSINT (see Section 2). Dalziel et al. [2] define CTI as: *Information that has been refined, analyzed or processed such that it is relevant, actionable and valuable with regards to an organization's security objectives.* In this context, the term is used to describe threat-related information which allows cyber security experts to investigate on a certain threat, e.g. the name of a malware, adversary or vulnerability. Additionally, it is considered actionable, if it is obtained in a timely manner meaning in due time to adapt the defensive measures to the threat in question before it hits in the form of an attack. Automating the collection of CTI can improve the defense capabilities against cyber threats but itself requires to face with the selection of the most relevant sources, the balancing between precision and timeliness that lead to an earlier generation of threat alerts, which in turn provides the security experts more time to prepare against potential upcoming attacks. Relying on the intelligence alone for an emerging threat is not assumed to be sufficient, and waiting for the occurrence of additional information to confirm the threat reduces the time advantage [14].

In this direction, this paper proposes an approach for the automated assessment of the OSINT sources themselves as an additional criterion for the relevance of CTI. In particular, an upstream assessment of the publishing source itself is taken into account, both when generating intelligence-based alerts and to decide whether a source should be used for CTI collection or not. In particular, a specific OSINT source has been selected based on a survey conducted among cyber security professionals and academic researchers who are working in the field of threat intelligence. Then two feature sets, that characterize the OSINT source have been defined. A scoring function to quantify the relevance of an OSINT source with regards to CTI in particular consideration of the timeliness has been proposed. The experimentation was conducted by training 5 regression models on both feature sets to predict the relevance score for OSINT sources, by focusing on Twitter, and compared with related approaches.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_34

The rest of the paper is organized as follows: in Section 2, the most related works are discussed. Section 3 elaborates the overall proposal in order to achieve the aforementioned objectives. The implementation details, the evaluation approach along with the gathered results are presented in Section 4, whereas Section 5 concludes this work.

2 Related Work

The growing interest in cyber threat intelligence (CTI) with regards to open sources (OSINT) is shown by the increasing research efforts in this field.

In [14], a ranking mechanism, to automate the evaluation of CTI sources and by selecting a subset of sources for CTI collection, is proposed. It deals with vulnerabilities disclosure in Twitter, by examining tweets which contain a Common Vulnerabilities and Exposures (CVE) ID. The authors showed that monitoring a subset of users on Twitter can be sufficient to retrieve most of the vulnerability-related information that is available on the microblogging platform. However, no ranking or scoring of the actual sources and their relevance is provided and they did not considered the detection of emerging malware and zero-day attacks. In [16], instead, the need for a quantitative evaluation of CTI sources is discussed and then an adaptive methodology for a weighted evaluation of such sources is proposed. The methodology introduces six evaluation categories on the basis of intelligence source aspects: (i) type of information, (ii) provider classification, (iii) licensing options, (iv) interoperability, (v) advanced API support and (vi) context applicability. The use of only structured data represents a limit in their methodology, furthermore, as the authors stated, other information such as the timeliness based on the time passed was not considered. In [6], a system, called Sec-Buzzer, for the detection of emerging topics related to cyber threats from expert communities on Twitter, is presented. It automatically identifies new experts on Twitter and adds them to a list of OSINT sources. In particular, the *activeness* of new candidates (i.e. potential experts) is evaluated on the basis of the number of tweets within a specified time period. The most active users are then further assessed according to their topic-relevance by examining the number of times they were mentioned in tweets and retweets by the most active existing experts. The main lack of this approach is that the user's activeness, as initial selection criterion, considers users with a high frequency of tweeting as experts. Even among cybersecurity-related Twitter accounts the number of tweets within a given time frame might not necessarily characterize a valuable threat intelligence source. In [15] instead, a system called DISCOVER is presented. It crawls both Twitter accounts of 69 international researchers and security analysts as well as a manually compiled list of 290 security blogs to discover emerging terms in the context of cyber threats. A natural language processing technique is used to preprocess the textual data as long as a list of terms related to emerging cyber threats is defined. They achieved 84% precision for warnings based on data from Twitter and 59% for the security blogs. Another research effort, called CyberTwitter [10], aimed to discover and analyze cybersecurity intelligence on Twitter,

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_34

collected in real-time by using Twitter API. The considered relevant information on cyber threats was then extracted on the basis of the Security Vulnerability Concept Extractor (SVCE), that is basically a Named Entity Recognizer (NER) specialized for cyber security terms. The automatic identification and generation of warnings was based on a set of properties, such as, the maximum time period for which intelligence is considered relevant. It showed that 57.2% of all inspected entities extracted by the SVCE were marked correctly and 33.2% were partially correct. From a total of 37 relevant intelligence entries the system generated 15 warnings, 13 of them were assessed as "useful" and the 2 remaining were "maybe useful". Then, 300 discarded tweets were manually examined by obtaining 85% recall. [11] extends [10] by introducing (i) National Vulnerability Databases (NVD), security blogs, Reddit and darkweb forums as additional OSINT sources along with Twitter, (ii) as well as a hybrid structure, called VKG, which combines knowledge graphs and word embeddings in a vector space. The approach was evaluated by manually annotating 60 alerts from which 49 were marked correct with a Precision of 81.6%. Furthermore, the SPARQL query engine was evaluated by searching for concepts that were marked "similar" by the annotators. Best results were reached for word embeddings with a dimensionality of 1500 and term frequency 2 which lead to a mean average precision of 69%. In [5], the authors tried to identify cyber threat-related tweets and gather CTI by linking mentioned vulnerabilities with their associated Common Vulnerabilities and Exposures (CVE). The proposed Centroid and the One-class Support Vector Machine (OCSVM) were compared to typical SVM, MLP, CNN, by showing that the Centroid novelty classifier using the cosine similarity distance performed better than the OCSVM with 85.1% Precision and 51.7% Recall. In [18], articles related to OSINT sources were examined, to gather insight into the semantics of malicious campaigns and the stages of malware distribution. The system extracts indicators of compromise (IOC) from security articles using regular expression, since they usually have fixed formats, e.g. IP address or hashsum. During the evaluation 91.9% Precision and 97.8% Recall for the IOC detection was reached and the stage classification through word embeddings resulted in an average Precision of 78.2%. In a survey reported in [17], emerged that cybersecurity experts are still unsatisfied with regard to the timeliness of many approaches that are currently used to collect CTI. The above presented research efforts and others [14][7] aimed to achieve earlier detection of cyber threats, by confirming the importance of such requirement.

From this review, important findings emerged, that were taken into consideration to narrow down the scope of this work. The main lack is due to the limited inspection to the textual data by neglecting the sources themselves for automated threat detection and warning generation. In light of such conducted analysis, the next Section elaborates the proposed approach in order to answer the following questions: (i) How to select relevant OSINT sources to be monitored, with high potential of publishing CTI, in order to avoid a large part of unreliable or outdated intelligence? (ii) How to automatically assessed the

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_34

threat intelligence's quality and credibility in order to issue a reliable warning for emerging threats?

3 Automated assessment of OSINT sources driven by features

In this Section, the process for automating the assessment of an OSINT source, for cyber threat intelligence, is described. The methodology, which is depicted in Fig. 1, can be organized in three main phases: *OSINT Sources Identification*, *Feature Selection* and *Score Definition*, that are elaborated in the following.

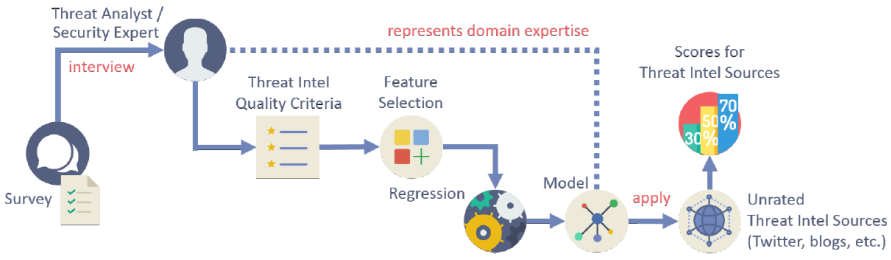

Fig. 1. Research method

3.1 OSINT Sources Identification

In the field of open-source intelligence a variety of public web sources, such as openly accessible web (e.g. vendor websites, Social network accounts, blogs) as well as forums and marketplaces in the darkweb, could be used to collect different types of threat intelligence. To deal with this challenge regarding the selection of relevant OSINT sources, an empirical study was conducted through an interview with 30 experts (ie. cyber security professionals and academic researchers) in the field of threat intelligence. The survey, which is used to establish the scope of this work but not the validity results, was based on the following questions:

1. What type of cyber threat intelligence is already being collected today?
2. How do experts rate the demand for improved CTI collection?
3. Which OSINT source are being utilized in today's CTI practice?
4. What are the most important criteria to be used to evaluate these sources?
5. How do experts rate certain sources with regards to their quality?
6. What features do the experts consider when evaluating the selected sources?

It aimed to retrieve information about (i) the type of CTI looked for in OSINT sources, such Zero-day vulnerabilities, CVE, IOC, upcoming malware, adversaries, etc. (ii) the characteristics to look for in a considered credible and qualitatively suitable source, such as technical details, code samples, author name,

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_34

outgoing links, google ranking, etc.; (iii) whether a set of OSINT sources are already being used or there are new one and how they would be rated with regard to quality, credibility; (iv) OSINT sources that are planned to be examined in the future or that might be worth to be examined by motivating that; (v) how often and how new OSINT sources are looked for, for example word of mouth, links found in specialized websites, search engines; (vi) how some provided CTI sources would be rated with regards to quality, credibility, TI domain and effort, when a manual searching and processing information is conducted.

Furthermore, the selected OSINT source types were quantified with regards to 4 different characteristics that are typical for threat intelligence, that is, (i) *Level of detail*: the source provides in-depth information about a threat, (ii) *Credibility*: the source provides credible intelligence (high true positive rate); (iii) *Timeliness*: the source provides intelligence in good time to act on it, (iv) *Actionable*: the source provides intelligence which can be used directly to support an organization's security objectives. Each criterion was rated on a scale from 0 (poor) to 5 (good) depending on whether the source usually provides intelligence with low or high quality for this criterion.

The first insight, according to the domain experts, was that the most important criteria for the evaluation of OSINT sources are both the *credibility* and the *timeliness* with which a source provides intelligence. The second one was that, among the top-5 types of cyber threat intelligence, as it is shown in Fig. 2-(a), 2 of them emerged (*vulnerability* and *malware*). In particular, by using the average value of the obtained values as the threshold, the demand for intelligence on "vulnerabilities and exploits" as well as "maleware" resulted particularly higher. In addition, the participants were asked to rate the most common OSINT source types from the related work: (i) public threat feeds, (ii) third-party websites and blogs, (iii) darkweb forums and marketplaces, (iv) Twitter, (v) Reddit, (vi) Pastebin and similar text & code storage websites, as it is depicted in Fig. 2-(b).

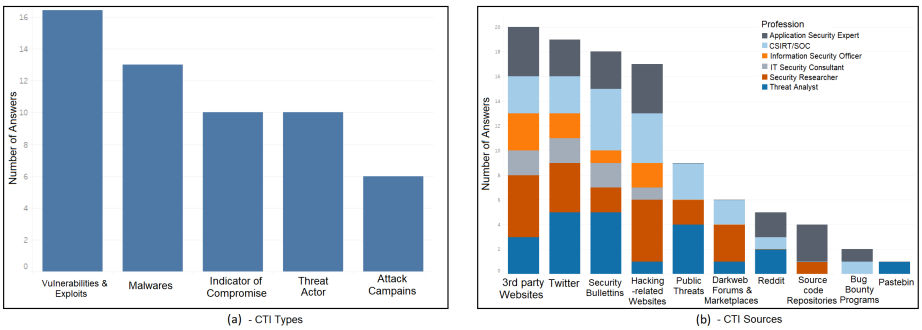

Fig. 2. CTI Types and CTI sources

Since the types (iii - v) comprise many sources (i.e. user accounts) for which the same meta data (i.e. features) is available, the experts were asked to select

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_34

the features that they considered promising or highly typical for valuable OSINT sources. They were also able to name additional features that they use when evaluating sources (see Section 3.2). On the basis of such insights, that third-party blogs, websites and Twitter emerged as the preferred sources for intelligence on new vulnerabilities and malwares. In particular, the CTI source was chosen by considering two main factors: (i) the popularity of the source in the context of threat intelligence, (ii) the type of available data that can be retrieved for supporting further analysis on them. Furthermore, even if Third-party website were rated higher with regards to the level of detail, Twitter is seen as a much more timely source type. Combining these findings and the fact that Twitter provides unified metadata on each user, which allows for better assessment and comparison, the author decided to investigate on Twitter as the OSINT source.

3.2 Feature Selection

From the analysis of related work, resulted that all existing methodologies aimed to identify cyber threat intelligence in different forms and qualities using natural language processing and machine learning techniques. Only few of them examined aspects of the source but none of them apply to the sources theirselves a feature-driven machine learning approach.

Table 1. Selected features based on related works and centered on Twitter meta-data

Feature	**Description**
num_mentions_community	The out-degree of the user in the mentions
num_hashtags	Total number of hashtags used in the observed time
ratio_retweets_replies	Ratio between retweets made by the user and replies received
num_mentioned_community	In-degree of the user in the mentions' monitored CTI social graph
num_retweets	Total number of retweets for a user
mean_mentions	Average number of mentions over all Tweets in the observed time period
num_tweets	Total number of tweets by a user
num_media	Total number of tweets containing media, for example images
verified	Whether the account has the 'verified' status by Twitter
num_likes	Total number of likes (favorites) received
num_following	Total number of friends, i.e. accounts that are followed by this user
days_since_join	Number of days since registration
mean_time_between_tweets	Average time between tweets during the observed time period in seconds
length_bio	Length of the user's description (biography)
mean_hashtags	Average number of hashtags per Tweet in the observed time period
num_followers	Total number of followers
length_username	Length of the displayed username
has_url	Whether the user profile has a website specified
length_url	Length of the website URL
mean_retweets	Average number of retweets made in the observed time period
num_mentions	Total number of mentions made by the user
mean_replies	Average number of replies received by the user
ratio_followers_following	Ratio between number of followers and following (friends)
mean_likes	Average number of likes (favorites) the user received
has_location	Whether the user profile has a location specified
num_replies	Total number of replies received by the user

On the other hand, various research efforts have been conducted to examine the role and characteristics of influencers on Twitter, such as, users who are

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_34

considered authoritative within a certain topical domain, as well as metrics to quantify the credibility of tweets and Twitter users. These approaches are often based on features extracted from profile meta data, the social graph and textual data from Tweets.

Based on such information, the first set of features, centered on meta-data listed in Table 1, has been selected by considering 3 aspects: (i) *Profile related features:* these are characteristics of a Twitter profile that are directly associated with the user profile (e.g. *registration date, the user's specified location, number offollowers* and so on). (ii) *Social graph related features:* this features are related to the connections (edges) among certain users (i.e. nodes) and allows to inspect the relations between them within a group or community of connected profiles. In particular: *followed/following, retweets* and *mentioned/mentions*, where in-degree and out-degree values of each node can be computed and compared. (iii) *Tweet related features:* other features, which are specifically associated to a single Tweet, that provide additional information on the user's behaviour with regards to the published Tweets. The second feature set is based on the *word embedding* technique, that is adopted to examine only textual content of the Tweets. It is based on "doc2vec" algorithm, with a 50-dimensional word embeddings as in [12], that strives when determining the similarity between different textual data.

3.3 Score Definition

In order to support the evaluation of the relevance of a threat intelligence source, a score function is proposed. It assigns a score R_I, between 0 and 1, to each threat intelligence source I on the basis of the weighted count of all true published intelligence $r_i \in I$. The proposed decay function, for calculating the score for a single CTI term r_i, is reprented through Equation (1).

$$r_i = score(t_i) = \begin{cases} 1 - 0.5 \cdot (\frac{t}{C^{-1}})^2 & s \cdot (c-1)^{1.25} \leq t \leq s \cdot c^{1.25},\ \ 0 < c \leq C,\ \ c \in N \\ 0.5 \cdot 0.5^{\frac{t}{s}} & s \cdot C < t \end{cases} \tag{1}$$

To include the timeliness of intelligence the weighting uses the time span that passed since a CTI term has been observed for the first time within the monitored community and the moment it is mentioned again by one of the other sources. In particular, this time delta t, which is determined in seconds, is then used as an input to the function which calculates the actual weight. Additionally, for a chosen number of intervals C the score is calculated as a step function such that slight time differences during the first few minutes or hours after the first occurrence of some threat intelligence do not influence the score. This was done because users considered intelligence sufficiently timely during an initial time period after the first occurrence and wanted a decrease in the score to indicate larger time differences, i.e. change in intervals. The value C = 5 has been empirically determined, and the size of the first interval was set to s = 86,400 which corresponds to the number of seconds in a full day. For intelligence which was observed exactly after the initial time intervals s*C, the

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_34

score is score(s*C) = 0.5 and intelligence mentioned later than this point of time gets a score below 0.5 assigned through the exponential decay function.

Then, all the r_i are aggregated per source I in order to assign a single relevance score to each source R_I according to Equation (2).

$$cti\_relevance\_score(R_I) = \frac{1}{|R_I|} \sum_{i=1}^{|R_I|} r_i \cdot \frac{log(|R_I|)}{log(|R|)} \tag{2}$$

In particular, the arithmetic mean is calculated over all single relevance scores $r_i = score(t_i)$ of a source $R_I = \{r_1, r_2, ..., r_I\}$ and weighted by the logarithmically normalized number of threat-related terms that were observed for this source, where R represents the full set of all scores and R_I the scores for intelligence shared by source I. After all sources are assigned a CTI Relevance Score, they are used to train a model to predict the relevance (see Section 4), measured through a value between [0,1], for other sources on the basis of their features.

4 Implementation and conducted experiments

In this Section, first data collection, the used regressor models and evaluation metrics are described and then the experimental results are reported.

4.1 Data collection, regression models and evaluation criteria

The data collection focused on Tweets and Twitter profiles, including metadata, related to the field of cybersecurity as a starting point to generate sets of training and testing data later on. In particular, an initial list of cyber security and cyber threat-related hashtags was manually compiled *(e.g. infosec, cybersecurity, security, threatintel, hacking, malware)*, as result of the survey.

This initial list of hashtags was then extended using the official Twitter API and third-party web services to find a more complete list of hashtags that are commonly being used in combination with one of the initial hashtags and therefore are assumed to be relevant to the field of cyber threat intelligence *(i.e. bugbounty, cve, cvss, cyberattack, cybercrime, cybercriminals, cybersec, databreach, dataleak, exploit, exploits, hacker, hackers, itsec, itsecurity, privacy, ransomware, redteam, threatintelligence, virus, vuln, vulnerabilities, vulnerability)*. This procedure was repeated on a daily basis from the 1st until the 31st of May 2019.

The official Twitter API was queried to retrieve the suggested hashtags listed under "Related Search" as well as three third-party web services, namely *keyhole.co*, *RiteKit* and *Hashtagify*. From each of these sources and for each of the hashtags in the current list, the top 3 hashtags, that is, those with the highest co-occurrence were retrieved and added to the list if they were not yet part of it. Each hashtag was used to also query the official Twitter API and retrieve the top 20 entries in the list of user accounts suggested by Twitter that recently used this hashtag. New suggestions were then added to a list of relevant Twitter users. After removal of duplicates 156 Twitter profiles remained and then they were merged with additional 16 Twitter profiles used in [5], by reaching a total of

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_34

172 profiles that represent the reference community on Twitter related to cyber threats and security. To be able to compare the features of users within this community against outside users that are not focused on cyber security, another list of Twitter profiles was retrieved from the Twitter API using the hashtags *technology, windows, linux, computer and internetofthings* while making sure that they were not in the list of suggested users of any of the cyber security related hashtags from above. This was done to ensure these users are related to the domain of technology and used similar vocabulary but are not focused on cyber threat intelligence. The full list of 230 Twitter users includes 172 (75%), who are considered the CTI community and 58 users from the technology domain who have no prominent relation to the cyber security domain. Finally, after the full list of sources was compiled the meta data of these 230 Twitter profiles as well as all 1,217,213 available Tweets from the time period of 3 years (from the 1st of Jan. 2016 until 31st of Dec. 2018) were collected using the official Twitter API.

Furthermore, 5 regression algorithms were evaluated and compared to the related works, by considering that no regression approaches were adopted in the context of CTI and for Twitter as an OSINT source. Specifically, the following one have been chosen as the no-regression version is typically applied in the relate work (i) SVM Regressor (SVR) by applying the Gaussian Radial Basis Function (RBF) kernel; (ii) Random Forest Regressor (RFR) have been used to establish a baseline for comparison; (iii) a Gradient Boosting Tree regression (GBTR) model, (iv) the Extra Trees Regressor(ETR), which is less susceptible to overfitting; and (v) a Multi-Layer Perceptron regressor (MLPR). The implementation of the regression models was based on "scikit-learn" Python library [1], and the configuration paramters are reported in Table 2.

Table 2. Description regarding the regression models configuration

Regressor	**Parameter configuration description**
SVR	The Gaussian Radial Basis Function (RBF) kernel has been used with the implementations default parameters, according to [12].
RFR, ETR	Maximum number of features considered for the best split is $\sqrt{(26)} \approx 5$, for the full source metadata, and $\sqrt{(50)} \approx 7$ for word embedding [3].
GBTR	500 boosting stages during training optimizing the least squares loss function and limiting the maximum depth to 5 nodes per tree, as in [4].
MLPR	Hidden layer size of 50 for the 50-dimensional word embedding features and a hidden layer size of 26 for the source meta data features, as in [12].

Whereas, the implemented evaluation criteria have been based on the following metrics, which are used to evaluate the performance of regression models: *Mean Squared Error (MSE)*: which is computed as the arithmetic mean of all squared errors that were made during prediction of a numeric value; (ii) *Coefficient of Determination (R^2)*: it represents the proportion of the variance in the dependent variable that is predictable from the features that the model was trained on [8]. It is used to assess how well a regression model fits the data set. In the following subsection, the results are presented and discussed.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_34

4.2 Results discussion

All five regression models were trained and evaluated on the collected data set, which was split into training and testing set, by using a 10-fold cross-validation strategy according to [14][18]. The experiments exploited the list of 659 CTI terms that were found across all 230 selected sources (i.e. Twitter accounts).

Fig. 3-(a) visualizes the [0, 1]-normalized scores for all sources that are sorted on the horizontal axis according to their true intelligence count (baseline). Whereas Fig. 3-(b) shows the Absolut Error between the real value and the predicted ones. It worth noticing that the MLPR performed worst of all models and its predictions have errors beyond the range of [-0.6, 0.6] and are therefore not depicted.

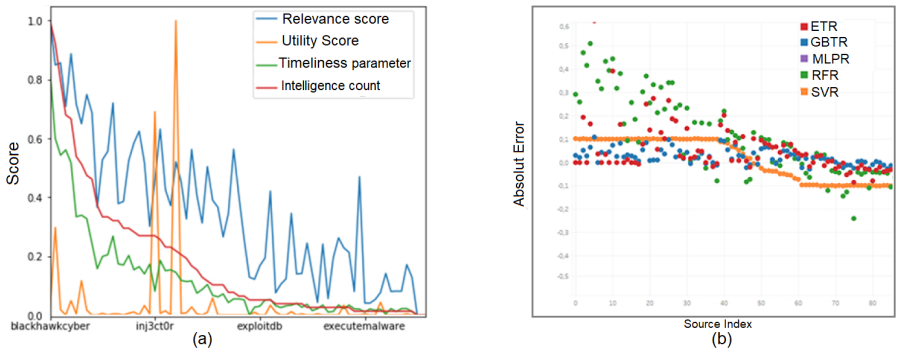

Fig. 3. CTI-Relevance-Score (a) and Absolut Error (b)

The R^2 value shows that best model for the prediction of the CTI Relevance Score on the source meta data feature set is the GBTR with an average value of $R^2 = 0.975$. The result evaluation is reported in Figure 4-(a). The same regression algorithms used for the source metadata feature set were trained on the word embedding model, that provides a single feature vector per Twitter source, by using identical parameters and metrics for training and evaluation. Figure 4-(b) displays a slight improvement in the R^2 for all models and even a large improvement for the MLPR model when using the word embedding features instead of the source meta data features. This first results indicate that the CTI Relevance Score can be predicted from CTI source features using the presented regression models.

The other question is about, whether such a score can be used to increase the timeliness of alert generation. Similar to the method described in [5], since each source can be represented by features derived from its metadata or a word embedding vector, both types of feature vectors were used to calculate three different centroids representing the community of CTI sources and the Tweets containing true intelligence, respectively: (i) Centroid based on the meta data features of all top sources from the CTI community, i.e. the top 30% of Twitter

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_34

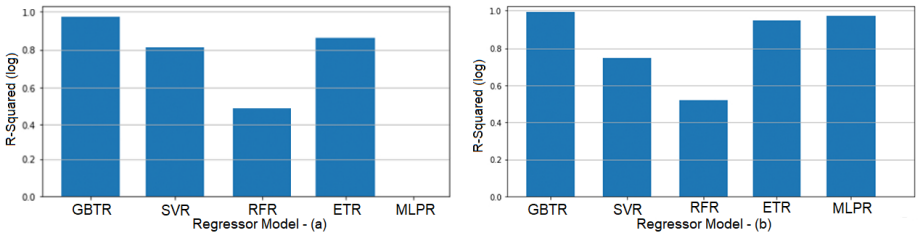

Fig. 4. Score prediction based on (a) meta-data feature set and (b) word embeddings

users with respect to their CTI Relevance Score; (ii) Centroid based on the word embeddings of all top sources selected analogous to the previous centroid; (iii) Centroid based on the word embeddings of all Tweets containing true intelligence not taking the source into account, to improve the identification of CTI Tweets.

The cosine-similarity between a source and the centroid is then interpreted as the score that quantifies the source relevance, i.e. a source similar to the community of already relevant CTI sources is thereby relevant as well. In order to determine if a CTI source is relevant a threshold t needs to be established such that only sources with a score above t are classified relevant. Fig. 5-(a) shows how the precision varies for possible thresholds t between [0, 1]. The red baseline indicates the precision $P_{base} = 45\%$ achieved on this data set using the count-based rule from DISCOVER [5] which only alerts on intelligence after their second occurrence. All scores reached higher precision for varying thresholds. The cosine-similarity to the third centroid (orange) reaches the highest precision but only for a rather high threshold which corresponds to a lower recall meaning that no alerts are issued for some intelligence. Considering a trade-off between a low threshold, i.e. high recall, and a high precision the F1-Score is calculated and showed in Fig. 5-(b).

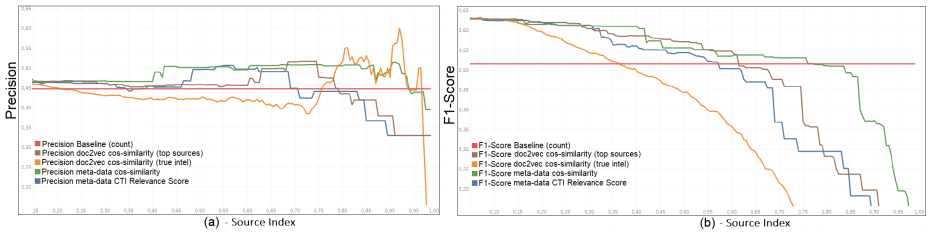

Fig. 5. Precision and F1-Score used to quantify the relevance of the predicted scores

This shows that the cosine-similarity to the third centroid (orange) is actually performing worse than all other scores. The cosine-similarity for the second centroid (brown) shows a slightly better F1-Score as the predicted CTI Relevance Score on the source meta data features (blue). Interestingly, the cosine-similarity

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_34

for the first centroid (green) has a F1-Score above the baseline for all thresholds up to t = 0.752. Through visual examination of the green graph a threshold of t = 0.4 is chosen to analyze the time advantage gained when using the cosine-similarity for the centroid of the source meta data features. This means that for any emerging CTI that is published by a source with a cosine-similarity above the selected threshold, an immediate alert is generated instead of waiting for a second occurrence of that intelligence from a different source. This time delay in hours is calculated for each instance in the dataset and visualized in Fig. 6. It shows not only the number of alerts that could be issued earlier but also the average time advantage to be gained: Half of all alerts could have been issued at least 32 hours earlier than other count-based systems like DISCOVER.

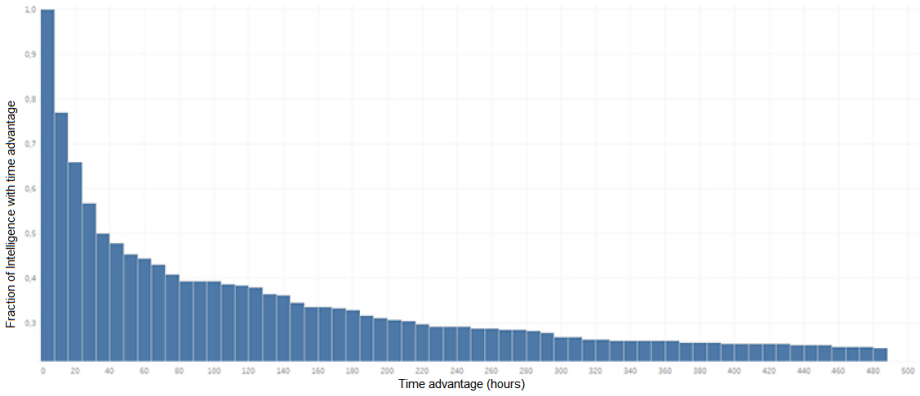

Fig. 6. The time advantage in hours gained when using the relevance score

5 Conclusion

This paper focused on the relevance assessment of OSINT sources as a cyber threat-related source. Two feature sets were engineered from the acquired data set and to quantify their relevance a CTI Relevance Score was formalized and compared with other scores. It emerged that the relevance of an open source on Twitter for CTI could be predicted through an automated feature-driven assessment of the source. As the results showed, half of all alerts could have been issued at least 32 hours earlier, meaning the time advantage of preventive cyber threat detection can be increased when using the quantified source relevance as a decisive factor for automated alert generation in existing systems.

Acknowledgment

This work was performed in the context of the CHAMPIONs project, which receives funding from the EU Internal Security Fund - Police, grant no. 823705.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_34

References

1. Scikit-learn: machine learning in python, https://scikit-learn.org/stable/
2. Dalziel, H., Olson, E., Carnall, J.: How to define and build an effective cyber threat intelligence capability. Syngress is an imprint of Elsevier, http://www.books24x7.com/marc.asp?bookid=78688, OCLC: 910537102
3. Devore, J.L.: Probability and statistics for engineering and the sciences. Brooks/Cole, Cengage Learning, eighth edition edn.
4. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30, pp. 3146–3154. Curran Associates, Inc.
5. Le, B.D., Wang, G., Nasim, M., Babar, A.: Gathering cyber threat intelligence from twitter using novelty classification http://arxiv.org/abs/1907.01755
6. Lee, K.C., Hsieh, C.H., Wei, L.J., Mao, C.H., Dai, J.H., Kuang, Y.T.: Sec-buzzer: cyber security emerging topic mining with open threat intelligence retrieval and timeline event annotation **21**(11), 2883–2896
7. Liao, X., Yuan, K., Wang, X., Li, Z., Xing, L., Beyah, R.: Acing the IOC game: Toward automatic discovery and analysis of open-source cyber threat intelligence. In: Proc. of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16. pp. 755–766
8. Marsland, S.: Machine learning: an algorithmic perspective. Chapman & Hall/CRC machine learning & pattern recognition series, CRC Press, second edition edn.
9. Mavroeidis, V., Bromander, S.: Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence. In: 2017 European Intelligence and Security Informatics Conference. pp. 91–98
10. Mittal, S., Das, P.K., Mulwad, V., Joshi, A., Finin, T.: 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. pp. 860–867
11. Mittal, S., Joshi, A., Finin, T.: Thinking, fast and slow: Combining vector spaces and knowledge graphs http://arxiv.org/abs/1708.03310
12. Nebot, V., Rangel, F., Berlanga, R., Rosso, P.: Identifying and classifying influencers in twitter only with textual information. In: Natural Language Processing and Information Systems. vol. 10859, pp. 28–39. Springer
13. Robertson, J.: Darkweb cyber threat intelligence mining. Cambridge University Press (2017)
14. Sabottke, C., Suciu, O., Dumitras, T.: Vulnerability disclosure in the age of social media: Exploiting twitter for predicting real-world exploits. In: 24th USENIX Security Symposium (USENIX Security 15). pp. 1041–1056
15. Sapienza, A., Ernala, S.K., Bessi, A., Lerman, K., Ferrara, E.: DISCOVER: Mining online chatter for emerging cyber threats. In: Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18. pp. 983–990
16. Schaberreiter, T., Kupfersberger, V., Rantos, K., Spyros, A., Papanikolaou, A., Ilioudis, C., Quirchmayr, G.: A quantitative evaluation of trust in the quality of cyber threat intelligence sources. In: Proc. of the 14th International Conference on Availability, Reliability and Security - ARES '19. pp. 1–10. ACM Press
17. Tounsi, W., Rais, H.: A survey on technical threat intelligence in the age of sophisticated cyber attacks **72**, 212–233 (2018)
18. Zhu, Z., Dumitras, T.: ChainSmith: Automatically learning the semantics of malicious campaigns by mining threat intelligence reports. In: IEEE European Symposium on Security and Privacy (EuroS&P). pp. 458–472 (2018)

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_34