Computers & Security 113 (2022) 102576

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

TC 11 Briefing Papers

A Feature-driven Method for Automating the Assessment of OSINT Cyber Threat Sources ☆

Andrea Tundis [a,*], Samuel Ruppert [b], Max Mühlhäuser [a]

[a] Technische Universität Darmstadt, Hochschulstrasse 10, Darmstadt 64289, Germany
[b] Deutsche Bahn AG, Frankfurt am Main, Germany

ARTICLE INFO

Article history:
Received 6 January 2021
Revised 10 November 2021
Accepted 9 December 2021
Available online 11 December 2021

Keywords:
Open source cyber threat intelligence
Cybersecurity
Machine learning
Feature engineering
Twitter

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. Currently, specific indicators of compromise (e.g. IP addresses, domains, hashsums of malicious files) are shared in a semi-automated and structured way via so-called threat feeds. Unfortunately, current systems have to deal with the trade-off between the timeliness of such an alert (i.e. warning at the first mention of a threat) and the need to wait for verification by other sources (i.e. warning after multiple sources have verified the threat). In addition, due to the increasing number of open sources, it is challenging to find the right balance between feasibility and costs in order to identify a relatively small subset of valuable sources. In this paper, a method 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.

1. Introduction

The use of heterogeneous technologies and IT networks has not only become essential in our everyday life but also comes at a price due to emerging vulnerabilities. Cyber Threat Intelligence (CTI) is a new field whose main mission is to research and analyze trends and technical developments related to Cybercrime, Hacktivism and Cyberespionage, based on the collection of intelligence using open source intelligence (OSINT), social media intelligence, human intelligence (Berghel, 2014). Some researchers are exploring OSINT as a means to proactively gather CTI from individuals and organizations that share relevant information publicly on the web (Robertson, 2017; Pastor-Galindo et al., 2020). Other efforts are related to malware analysis and intrusion detection systems for threat intelligence on a technical level, for example by generating specific indicators of compromise (IOC) in a machine-readable format. This offers a high potential for CTI with regards to the atomic indicators or host and network artifacts from a detected attack (Mavroeidis and Bromander, 2017), however it is mostly limited to *situational awareness* and *imminent threats* towards a specific organization (Robertson, 2017). Indeed, the current approaches are not designed to provide threat analysts with a broader "understanding of an adversary's capabilities" or even general "knowledge about the adversaries themselves". This scenario and as timeliness is essential in security emphasize the need to determine the relevance of such information not only based on whether it is widely spread but also on the quality and informativeness of the source itself (Tounsi and Rais, 2018).

On the one hand there is a number of software or hardware vendors aiming to inform their customers about current threats to-

☆ This is an extended version of the paper "Tundis A., Ruppert S., Mühlhäuser M. On the Automated Assessment of Open-Source Cyber Threat Intelligence Sources. In: Krzhizhanovskaya V. et al. (eds) Computational Science (ICCS 2020). Lecture Notes in Computer Science, vol 12138. Springer, Cham."

* Corresponding author.
E-mail addresses: tundis@tk.tu-darmstadt.de (A. Tundis), samuel.ruppert@deutschebahn.com (S. Ruppert), max@tk.tu-darmstadt.de (M. Mühlhäuser).

https://doi.org/10.1016/j.cose.2021.102576
0167-4048/

wards their products, e.g. newly found vulnerabilities, through an official bulletin, website or the like. Different publishers provide cyber threat-related information on vulnerabilities, adversaries or imminent attack campaigns in the Web. Many cyber security professionals, vendors and researchers announce their findings on business or personal websites, blogs and social media, which can be assumed to have the highest quality in terms of credibility and technical detail. However, even some hackers post information about ongoing attack campaigns or new vulnerabilities on social media like Twitter and on some forums and marketplaces in the darkweb. This information might often reach cyber security professionals rather late and to be unprepared for such attacks (Sabottke et al., 2015). 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. Nonetheless, these sources have already proven to hold valuable additional information for an organization's defense against cyber attacks (Tounsi and Rais, 2018). To understand these evolving threats, it is essential for security experts to illuminate the threat landscape including adversaries, their tools and techniques (Mavroeidis and Bromander, 2017), as well as learn about the details of cyber threats relevant sources and prioritize them (Robertson, 2017). It is simply not practical to implement counter-measures in a timely and economical manner for all possible attacks including 0-day exploits, i.e. exploits that target vulnerabilities that the defenders do not know about it yet (Stone, 2021). For this reason, automating the collection of CTI (Dalziel et al., 2015) (i.e. publicly accessible data on the internet) can represent a viable approach to improve the defense capabilities against cyber threats, but of course it 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.

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 was 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 were defined. A scoring function to quantify the relevance of an OSINT source with regards to CTI by particularly considering the timeliness was 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. The rest of the paper is organized as follows: in Section 2, the most related works are discussed, whereas further motivations are presented in Section 3. Section 4 elaborates the overall proposal; whereas, the implementation details, the evaluation approach and the gathered results are presented in Section 5. Finally, Section 6 concludes this work.

2. Related work

This section discusses the most relevant related works organized by topic, in cyber threat intelligence (CTI) with regards to OSINT, whose main aspects are summarized in Table 1.

Vulnerability Disclosure in Social Media The paper in Sabottke et al. (2015) dealt with vulnerabilities disclosure in social media by examining tweets which contain a Common Vulnerabilities and Exposures (CVE) ID, which can be used to classify the vulnerabilities according to their exploitability and reach a more realistic measure for real-world exploits than the Common Vulnerability Scoring System (CVSS) score. It proposes a ranking mechanism, based on supervised machine learning, to automate the evaluation of open-source CTI sources by selecting a subset of sources for CTI collection. It has been 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. The authors discussed the growing rate of vulnerability discovery over the years a well as their publication through coordinated disclosure processes which often leads to a flood of vulnerability disclosures within a narrow time frame. However, this raised the challenge to select such users, i.e. sources, on Twitter depending on their informativeness and relevance for early exploit detection by proposing a new measurement to quantified each user's utility based on their number of relevant tweets related to real-world exploits. Moreover, even it is an important ground for the scope of this current work, (i) no ranking or scoring of the actual sources and their relevance was provided and, (ii) they did not considered the detection of emerging malware and 0-day attacks.

Evaluation of Trust in the Quality of CTI Sources In Schaberreiter et al. (2019), the need for a quantitative evaluation of CTI sources is discussed and then an adaptive methodology for a weighted evaluation of such sources was proposed by using Structured Threat Information Expression (STIX) OASIS Open, that is a language for describing cyber threat information, which is used to exchange CTI in a structured manner. 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. Each category was evaluated through multiple parameters, represented as numeric values, which described the quality of the CTI within a closed world view. This means that all sources were only compared against other sources within the monitored community of CTI sources. Unfortunately, the use of only structured data (i.e. data fields of the STIX format) represents a big limit of their methodology, furthermore, as the same authors stated, other information such as the timeliness based on the time passed was not considered.

Sec-Buzzer A system for the detection of emerging topics related to cyber threats from expert communities on Twitter, called Sec-Buzzer, was presented in Lee et al. (2017). It automatically identifies new experts on Twitter and adds them to a list of OSINT sources. The most active users in posting information 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 remaining candidates are then filtered using the number of their tweets that contain keywords from a predefined list of relevant topics. These users are then proposed as additional expert candidates and manually confirmed. Then, all expert-users were weighted according to two hypotheses: (i) if expert A is followed by another expert B with a higher weight, the weight of expert A increases, (ii) the more attention an expert can attract, the higher is the weight for this expert. 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. However, even among cybersecurity-related Twitter accounts the number of tweets within a given time frame might not necessarily characterize a valuable threat intelligence source.

Early Warning of Cyber Threats In Sapienza et al. (2017a), Twitter textual data is explored to generate alerts on emerging cyber threats like DDoS attacks and data breaches. The system tries to confirm any relevant keywords obtained from Twitter data by crawling a selection of darkweb forums and marketplaces. New threat-related terms are extracted by filtering common words,

Table 1
Overview of the approaches related to the acquisition of cyber threat intelligence from open sources.

RW	Data Source	Type of Intelligence	Features & Implementation	Rules & Constrains	Alert Evaluation
Sabottke et al. (2015)	Twitter	Vulnerabilities (CVE)	Source Metadata, Textual data, linear SVM	Utility Score (Informativeness)	No evaluation for utility score
Schaberreiter et al. (2019)	MISP	STIX, OpenIOC	Semantic, Syntactical features of STIX objects	Timeliness parameter	None
Lee et al. (2017)	Twitter, Security Blogs	Mid-level CIT terms (TTP)	Source Metadata, Social graph, Textual data	Expert Authority weighting	Precision (81.4%), Recall (~20%)
Sapienza et al. (2017a)	Twitter	Mid-level CIT terms (TTP)	Keyword filtering, Term frequency	Rule set (Intelligence count)	Case study
Sapienza et al. (2018)	Twitter, Darkweb	Mid-level CIT terms (TTP)	Keyword filtering, Term frequency	Rule set (Intelligence count)	Case study, Precision (84%)
Mittal et al. (2016a)	Twitter	Cyber threat topic & concepts (TTP)	Keyword filtering, Term frequency, Tagging NER	SWRL rules (source count)	Precision (57%), Recall (~40%)
Mittal et al. (2017c)	Twitter	Cyber threat topic & concepts (TTP)	Keyword filtering, Tagging NER, Word embeddings	SPARQL/SWRL, System profiles	Precision (78%)
Le et al. (2019)	Twitter	Vulnerabilities (CVE)	Source Meta data, Term Frequency, Centroids	Cosine similarity	Precision (85%),Recall (52%), F1-Score (64%)

symbols and stopwords. Any token that occurs in the context of threat-related words, such as "botnet", is considered potential threat intelligence. If such tokens can also be observed in hacking-related online discussions in the darkweb, they are considered relevant CTI, thus reporting them as alerts. All alerts were manually annotated to determine the overall precision of the system, based on the 661 alerts issued on a period of 5 months with 83.1% precision. One case showed that the "Mirai" botnet-based DDoS attack was mentioned on Twitter more than 1 month before the attack happened, and 3 weeks before exploits were shared on the darkweb, by highlighting the potential of CTI from OSINT sources as Twitter.

DISCOVER In Sapienza et al. (2018) a system called DISCOVER, based on Sapienza et al. (2017a) and centered on natural language processing (NLP), 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. The data is stored in a database which is queried to generate warnings according to 2 main rules: (i) the novel term must have occurred more than once across all data sources, and (ii) at least one keyword from a threat-specific dictionary must be present in the same data. These rules were used to prevent a high false positive rate and thereby they achieved 84% precision for warnings based on data from Twitter. Five case studies were conducted to evaluate the system. They showed that DISCOVER issued a warning for the new term "wannacry" on April 18, 2017 which was about 3 weeks in advance to the wide-spread outbreak of the Wannacry ransomware. Though the system identified mentions of this term even before, a warning was not generated due to the required number of occurrences to be considered relevant. Moreover, how to allow an earlier generation of warnings while maintaining high precision was not investigated.

CyberTwitter In Mittal et al. (2016a), CyberTwitter, which aimed to discover and analyze cybersecurity intelligence on Twitter, collected in real-time, is presented. The considered relevant information on cyber threats was extracted on the basis of the Security Vulnerability Concept Extractor (SVCE). The relations between these entities are described using a Unified Cybersecurity Ontology (UCO) and stored in a Cybersecurity Knowledge Base (KB). Any reoccurring intelligence was used to update the database which stores a number of properties on each entry. 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. Additionally, and through the use of a rule set written in Semantic Web Rule Language (SWRL), custom user system profiles could be configured by defining, for example, for which software and hardware some warnings should be generated. The approach was evaluated through human assessment by doctoral students in the field of cybersecurity and using data collected over a ten-day time frame. 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.

Cyber-All-Intel Mittal et al. (2017c) extends Mittal et al. (2016a) by introducing (i) National Vulnerability Databases (NVD), security blogs, Reddit and darkweb forums as additional OSINT sources along with Twitter, as well as (ii) a hybrid structure, called VKG, which combines knowledge graphs and word embeddings in a vector space. To derive threat intelligence and insights from the VKG, the SPARQL query language was extended with a layer that integrates vector embeddings, whereas, alerts were generated by using the rule set defined in Mittal et al. (2016a). The approach was evaluated by manually annotating 60 alerts from which 49 were marked correct. 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. Overall, the vector embeddings performed better than the knowledge graph when searching for similar attacks, products and vulnerabilities. Out of the 55 alerts issued by the system for the evaluation, 43 were manually marked "useful" and 9 as "maybe useful" which a precision of 78%.

Gathering CTI using Novelty Classification In Le et al. (2019), the authors tried to identify cyber threat-related tweets and gather CTI by linking mentioned vulnerabilities with their associated Common Vulnerabilities and Exposures (CVE).From a set of 50 cybersecurity related users a set of features were extracted and used to train a classifier of "novelty". Each tweet and CVE description was converted into a numerical vector representation where each feature is the Term Frequency-Invert Document Frequency (TF-IDF) for all document terms. Afterwards, the Centroid and the One-class Support Vector Machine (OCSVM) novelty classifiers were trained using these features. Both classifiers were compared to typical SVM, MLP, CNN, on a data set collected over 12 months in 2018, by showing that the centroid novelty classifier, using the cosine similarity distance, performed slightly better than the OCSVM with 85% Precision and 52% Recall, from a total of 232 cyber threat-related tweets. From 81 tweets containing threat intelligence without a

CVE mentioned, 34 could be correctly linked to a CVE description based on such features and on the cosine similarity.

ChainSmith In Zhu and Dumitras (2018), different articles related to OSINT sources were examined, to gather insight into the semantics of malicious campaigns and the stages of malware distribution. Typically, the collection of indicators of compromise (IOC) from security articles would mean that the intelligence might already be outdated due to the time it takes from the occurrence of a threat until the publication of the corresponding security article. However, this approach does not aim to utilize these IOC in day-to-day security operations but rather to gather general insight into the stages of malware campaigns. Therefore, a word embedding algorithm was used to classify all IOCs according to four pre-defined stages of such campaigns (i.e. baiting, exploitation, installation, command & control). 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%. From such experiments, insights into some attack patterns and strategies applied by threat actors, which corresponds to the high-level intelligence of the CTI model, were gathered.

Other research contributions that dealt with other aspects, within threat intelligence field, are available in the literature. For example, in Bouwman et al. (2020) the use of threat intelligence from a commercial point of view, and in particular regarding paid threat intelligence, has been considered. Then, how to characterize in a more formal way different types of public and commercial threat intelligence sources has been addressed in Li et al. (2019), where a comparative analysis has been conducted by proposing a set of metrics. Whereas in Chen et al. (2019), the authors investigated the problem of predicting when an exploit is first seen, by proposing a framework to support the decision process of allocating resources in order to take corrective actions.

The above research efforts show the increasing interest in automate the collection of CTI from different OSINT sources and their use to generate alerts on emerging threats. However, only few of them investigated specific characteristics of the CTI publishing OSINT source itself. For example, in Sabottke et al. (2015) the Utility Score was introduced as a specialized measure of informativeness of CTI sources with regards to exploitable vulnerabilities, whereas, in Lee et al. (2017), a Utility Score as well as Expert Authority Weighting were both used to narrow down a set of CTI sources which are considered the most promising for the collection of CTI to be monitored. Nevertheless, these measures are based on the intelligence already published by a source in the past and do not aim to predict the relevance of new sources. In Schaberreiter et al. (2019), the proposed methodology focused on well-formatted CTI using a standardized syntax. Furthermore, systems like “DISCOVER” (Sapienza et al., 2018) or “Cyber-All- Intel” (Mittal et al., 2017c) cannot utilize this methodology to its full extent, since they aimed to detect very early topics and terms related to cyber threats and they cannot get all the required data from unstructured OSINT.

The discussed works (Lee et al., 2017; Sapienza et al., 2017a; Sapienza et al., 2018; Mittal et al., 2016a; Mittal et al., 2017c) based their evaluation mostly on the precision and recall of the CTI collection. The generation of alerts was usually evaluated in a qualitative manner, such as in (Sapienza et al., 2018), which limits the direct comparison of different rules used for alert generation. The datasets used for evaluation in (Zhu and Dumitras, 2018; Lee et al., 2017; Sapienza et al., 2017a; Sapienza et al., 2018; Mittal et al., 2016a; Mittal et al., 2017c) were all manually annotated for training and evaluation which leads to rather small datasets.

Additionally, the recall of such approaches can only be estimated for a real-world application with high uncertainty due to the small data sets used. Moreover, the rules for alert generation only take basic data about the intelligence into consideration, for example the time it was published and whether it contains predefined keywords. Information about the source from which the intelligence originated is not used in any of these systems, which limits the potential of early cyber threat detection.

In a survey reported in Tounsi and Rais (2018), it 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 Sabottke et al. (2015) and Liao et al. (2016) aimed to achieve earlier detection of cyber threats, by confirming the importance of such requirement. However, these approaches inspect only the textual data and do not evaluate the expertise or reliability of the sources. Regarding social media users, for example Twitter accounts, two related approaches have been researched: (i) assessing the credibility of tweets about live events like natural disasters (Krzysztof et al., 2015; Yang et al., 2019; Hassan, 2018; Khodabakhsh et al., 2018) and (ii) the detection of influencers who are often considered to be experts for a specific topic (Nebot et al., 2018; Subbian and Melville, 2011; Lahuerta-Otero and Cordero-Gutiérrez, 2016; Chu and Kim, 2011). Such related works have a different scope, however, they still provide relevant findings for feature engineering and selection with regards to OSINT sources on social media and especially Twitter.

Most of the presented related works share similarities regarding the objectives, implementation and evaluation procedures, however they neither share a common threat-specific dictionary for their textual pre-processing steps nor use the same data set for training and evaluation. This means that a direct comparison of the results, performances, and solutions also with our work is only possible to a limited extent. This is due to (i) the lack of standardized or sufficiently extensive data sets publicly available for cyber threat intelligence from OSINT sources (Bridges et al., 2017), and (ii) the methods used to acquire such data sets are not described in detail in the relted works and cannot be reproduced. An exception to this is (Le et al., 2019) where the full list of Twitter users as well as the time period during which their tweets were collected are explicitly stated, even if without any explanation about the section criteria of the users.

In conclusion, some important findings emerged, that were used 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, as further highlighted in the following section.

3. Motivations, background and challenges

Different types of CTIs are usually being utilized on different levels of an organization’s defenses (e.g. to create new firewall rules, to inspect network traffic and prioritize security patching), which are formalized in Mavroeidis and Bromander (2017). Specifically, low-level threat intelligence categories (i.e. “Target”, “Atomic Indicators” and “Indicators of Compromise”) are summarized as indicators of compromise (IOC), such as IP addresses, domains, hashsums of malicious files etc., that are known to presented threats in the past or present. Such intelligence is often easier to obtain, share and process in an automated manner, since it is the most structured intelligence with a fixed CTI format or syntax. The three lower levels of CTI are then followed by the mid-level categories (i.e.“Tools” and “TTPs”) that stand for tactics, techniques and procedures of a threat actor. They can sometimes overlap since the tools used by attackers could fall into the category “TTPs”, that is why they are summarized as TTP in this work; whereas high-level CTI categories are mostly of a strategic nature and only marginally relevant to this work.

Existing proof-of-concepts, as described in Section 2, have also shown that these sources have a potential to further improve insights into the cyber threat landscape which is not being fully utilized to date. For example, Sapienza et al. (2018) presented a case study explaining that their OSINT system detected mentions of the "wannacry" ransomware on Twitter about three weeks before the wide-spread infection of more than 200,000 computer systems across 150 countries. An automated warning at such an early point in time would most likely have been sufficient for many organizations to protect their networks against the spread of this malware. However, no warning was generated at that time since the system was designed to wait for the occurrence of additional, confirming intelligence in order to maintain an acceptable false positive rate.

Typical challenges for such an acquisition of CTI arise from the nature of the user-generated content that OSINT relies on. The less structured data normally lacks consistent metadata or syntax which requires at least basic techniques of natural language processing to even identify threat intelligence. Especially mid-level intelligence, (i.e. emerging TTP and malwares), which is less structured and has no specific format, is thereby harder to gather from large textual data sets like social media content. For example, the names of new malware families are not automatically obtained by intrusion detection systems, but rather assigned by security researchers. Narrowing the search for intelligence down to such sources that have a high probability of publishing relevant information is a non-trivial, yet important initial step to reduce noise in the data and simplify data cleaning considerably. In fact, both the quantity and the quality of the valuable information depends on the type of the selected CTIs and the OSINT sources from which they come. Many proof-of-concept systems introduced in recent research either focus only on a very limited subset of available sources for their evaluation or struggle to achieve near real-time monitoring of the OSINT sources due to the high amount of data. Such approaches might be difficult to apply in real-world scenarios unless this issue can be overcome without accidentally excluding important sources from the scope of such a system and thereby reducing its recall. For the collection of vulnerability-related CTI from different sources on Twitter (i.e. micro-blogs), it could be shown that monitoring only a manageable subset of sources is sufficient to obtain relevant information, as lot of intelligence found on Twitter is spread via retweets and mentions but, it has its origin in a relatively small subset of sources (Sabottke et al., 2015; Tundis and Mühlhäuser, 2017; Tundis et al., 2018).

Identifying threat intelligence from OSINT sources is mostly achieved by processing the unstructured text before classification algorithms (Sabottke et al., 2015; Zhu and Dumitras, 2018; Long et al., 2019; Le et al., 2019), or rule sets are used to decide whether it is considered real intelligence (Sapienza et al., 2018; Sapienza et al., 2017a; Mittal et al., 2017c; Mittal et al., 2016a). When automatically generating alerts based on this intelligence, it is crucial that the false positive rate is kept low in order to prevent flooding the cyber security analysts with alerts, and let them keep focus on the most critical threats. Unfortunately, such verification is challenging since there is often very limited information about cyber threats, e.g. the name of a new malware. Waiting for the occurrence of additional, confirming intelligence from other sources reduces the false positive rate, but it sacrifices the advantage in terms of the detection time. This trade-off concerns all approaches known to the authors and reviewed in Section 2, which represents one of the primary challenges in this field as it also emerged from a survey conducted among cyber security professionals in 2018 Tounsi and Rais (2018).

To grasp these constantly evolving threats, it is essential to understand the threat landscape including different adversaries, their tools and techniques (Mavroeidis and Bromander, 2017). However, it is challenging to cope with cyber threats even with forefront defensive measures like intrusion detection systems and modern firewalls due to different aspects to deal with, such as:

- *Understanding* - it is simply not practical to implement countermeasures in a timely and economical manner for all possible attacks including 0-day exploits. Thus, learning about the details of cyber threats, which are relevant within its own application domain and prioritizing them, is a vital step in defending computer systems.
- *Acquisition* - the acquisition of CTI arises from the nature of the user-generated content that OSINT relies on. The less structured data normally lacks of consistent metadata or syntax, thus requires at least basic natural language processing techniques to even identify threat intelligence.
- *Verification* - relevance assessment of emerging threats is hard due to limited information. Relying on the intelligence alone for an emerging threat is insufficient, and waiting for the occurrence of additional information to confirm the threat reduces the time advantage. As a consequence, it is necessary to find a trade-off to cope with it.

Automating the collection of CTI from OSINT sources can improve an organization's defense capabilities against cyber threats but itself requires to face with the aforementioned challenges in order to increase its benefits in real cases. Selecting the most relevant sources and balancing the trade-off between precision and timeliness might lead to an even earlier generation of threat alerts giving the security experts more time to prepare against potential attacks while maintaining the precision achieved by today's existing systems.

On the best of our knowledge, an upstream assessment of the publishing source itself, to be taken into account both when generating intelligence-based alerts and to decide whether a source should be used for CTI collection or not, has not been considered yet. We believe that the publishing source itself needs to be assessed and taken into account when generating intelligence-based alerts. As a consequence, Section 4 elaborates our proposal by facing with 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 assess the threat intelligence's quality and credibility in order to issue a reliable warning for emerging threats?

4. Automating the assessment of open-source cyber threat sources

In this Section, the overall approach for automating the assessment of an OSINT source, for cyber threat intelligence, is presented. At first, a general overview of the adopted research methodology is described in subsection 4.1. Then, its 4 main macro phases, called *OSINT Sources Identification, Feature Selection, Score Definition*, and *Model Training and Results Evaluation* are elaborated in greater detail in the next sections.

4.1. Research methodology

Fig. 1 depicts the general workflow that has been adopted. In the first phase, called *OSINT Sources Identification*, the problem related to the identification and the choice of the OSINT source, towards the focus has to be narrowed, has been faced. Starting from the chosen OSINT source, the second phase, called *Feature Selection*, dealt with the selection of its most significant metadata to be used. That is to say, not only to identify which features can be obtained from it, but first of all which are the most useful one, that better characterize the source on the basis of the study under consideration. The third phase, called *Score Definition*, focused on how

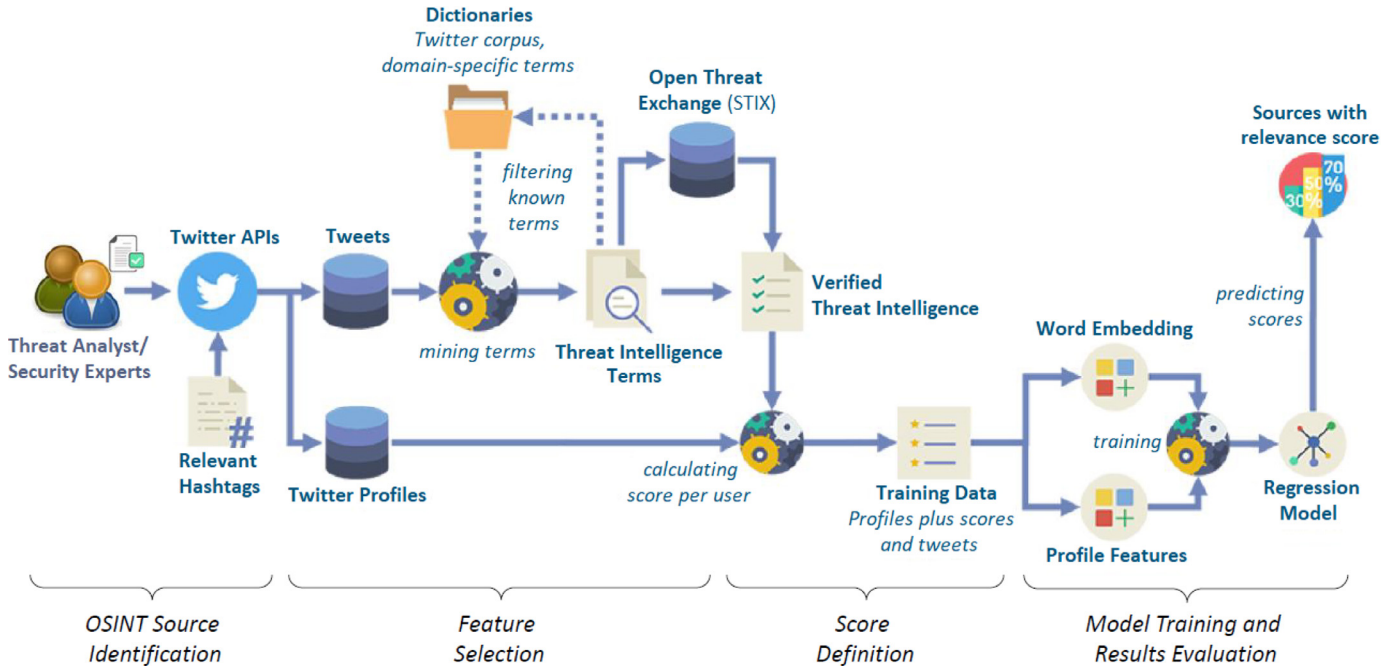

Fig. 1. Research method overview.

to use these features, and their related values, in order to provide an assessment score of each source. Finally, the last phase, presented in Section 5, consisted of the concrete implementation and experimentation of the method, centered on the use of different regression models.

It is worth noting that, whilst previous research efforts, such that those discussed in the related work Section, such as (Sapienza et al., 2018; Le et al., 2019; Mittal et al., 2016a), dealt with single pieces of intelligence separately, this method faces with intelligence-related info in conjunction.

4.2. 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.

Some OSINT sources provide intelligence on low-level CTI, e.g. malicious IP addresses, which is usually originates from the network or endpoint sensors like intrusion detection systems or antivirus software and it is often shared through automated platforms on the web. By contrast the mid-level intelligence like vulnerabilities, exploits or malware names typically originates from manual analysis by security researchers or other security experts which means that it is likely to be published from different OSINT sources on the web. Such intelligence can yield a high time advantage and allow further investigation into emerging threats even before attacks happen and network sensors detect them. However, not only this requires the monitoring of the right OSINT sources but also to apply specific method of filtering or verifying the intelligence before alerts are generated and displayed to the human threat analyst. It can be safely assumed that each source type has its own characteristics along with certain advantages and drawbacks when it comes to real-world requirements and security objectives. Since no in-depth comparison of the different source types for CTI exists to the knowledge of the authors, the question arises which source types are preferred for mid-level CTI in real-world applications.

In order to answer this question and thereby direct this work towards the most promising OSINT source types and criteria, an empirical study was conducted through an interview with 30 experts (i.e. security professionals in the industry and academic security researchers) in the field of threat intelligence. They were recruited through direct contacts and by searching online services, like (e.g. LinkedIn) and double checking on the web page of their organization, for relevant keywords (e.g. cyber security, threat intelligence, etc.). Only participants with dedicated IT security job description and a minimum of 5 years of experience in this field were contacted and invited to participate in the online survey. While not all of them made regular use of OSINT CTI as part of their daily work, as it is emerged from the survey described in the following (e.g. application security experts), many of them expressed that they were planning to utilize CTI and already had concrete expectations, thus contributing an additional perspective to the survey. Based on that, the overall goal of such survey was to provide a better insight into the cyber threat intelligence domain and determine which OSINT source types are currently of interest for the domain experts and what characteristics are considered during the evaluation when deciding which sources are worth monitoring. The survey was based on the following questions:

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

As one can see, these and many other interesting questions can be asked in such research field. However, only the first four questions have been deeper tackled in this paper, thus obtaining the necessary information, in order to achieve the main objective within the current research context. Whereas, questions 5 and 6 could be considered more in deep for future investigations.

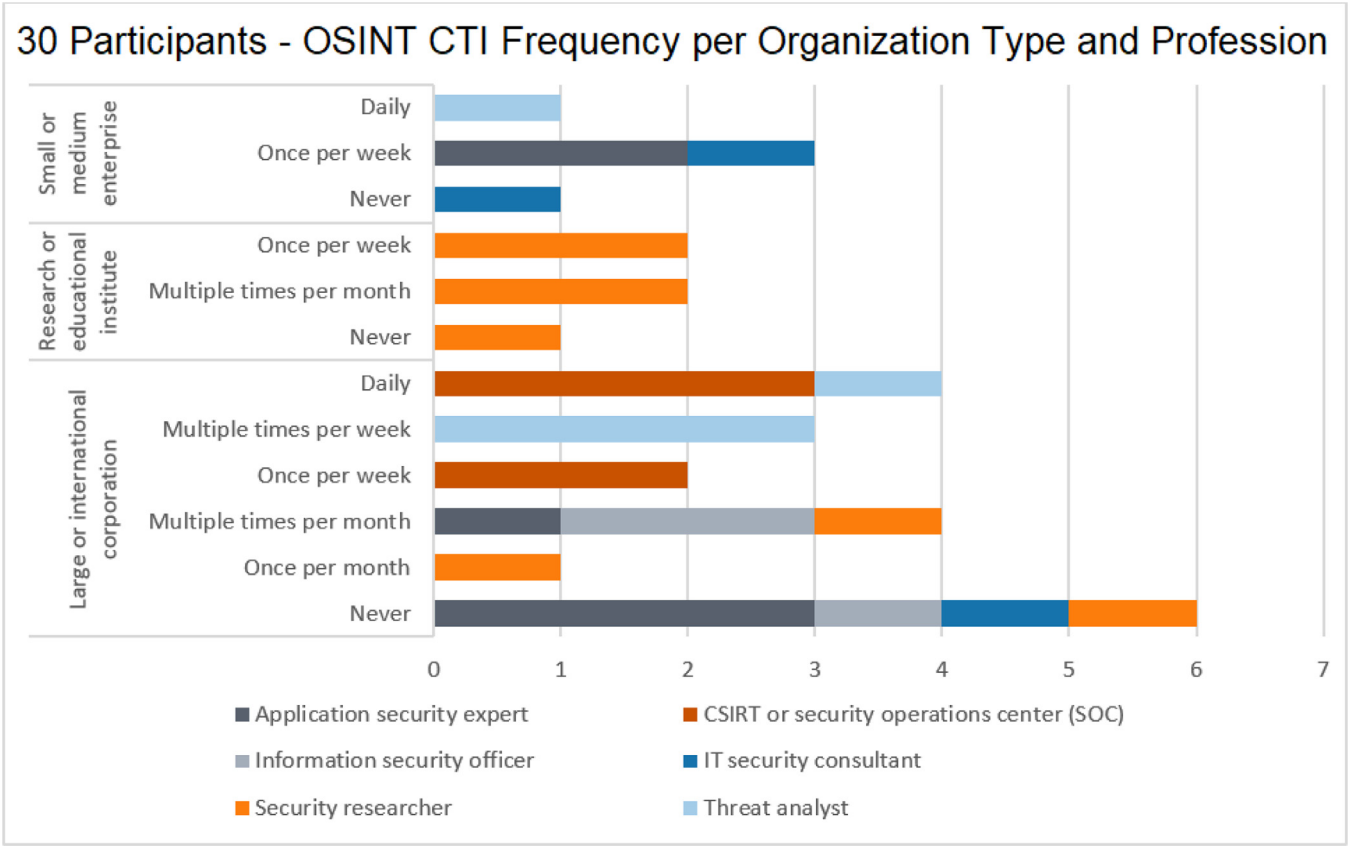

Fig. 2. Cyber security experts, professions and frequency of OSINT usage for CTI.

This survey 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, 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 characteristic (i.e. *Level of detail, Credibility, Timeliness*, and *Actionable*) was rated, from the experts, on a scale from 0 (poor) to 5 (good) depending on whether the source usually provides intelligence with low or high quality for such characteristic.

Fig. 2 shows the 30 participants of the survey representing six different professions, that is *Threat Analyst, Security Researcher, Application Security Expert, IT Security Consultant, CSIRT/Security Operations Center (SOC), Information Security Officer* of the cyber security field, by highlighting how frequently the participants of the survey utilize open-source intelligence (OSINT) for CTI as part of their regular work. One can see that, the most frequent use of cyber threat intelligence is made by *Threat Analysts* and personnel of *Security Operations Centers*.

Around 75% of the participants stated that they use OSINT for CTI during their work, whereas around 25% of them said that they have never used it, which could be a notably percentage. Especially *Application Security Experts* and *Information Security Officers* in large corporations rarely or even never use OSINT for CTI, since such organizations often have dedicated personnel for CTI analysis or focus on non-OSINT CTI such as commercial threat feeds. However, all of them considered CTI beneficial to their work based on their personal experience and confirmed the benefits which could derive from it. In fact, in the survey, none of them used neither the "I'm not sure" option nor left the question unanswered. With that, we want to emphasize that, this statement refers to the question regarding CTI types that are being considered beneficial by the participants and it is related to the lead questions five and six, that have been above listed. It worth noting that, the results are shown in Fig. 4 with the option "I'm not sure" being omitted since it was not chosen by any of the participants.

The first insight, according to the domain experts, who could express a single preference, is represented in Fig. 3. It shows that the most important criteria for the evaluation of OSINT sources are both the *credibility* and the *timeliness* with which a source provides intelligence.

As described in Section 3 different types of CTI are being used on different levels (i.e. operational, tactical, strategic). In this case, the participants of the survey were asked to indicate two CTI types, that are considered the most important one within their work. As it is shown in Fig. 4, by using the average value of the obtained values (by using checkboxes) as the threshold, it emerged that,

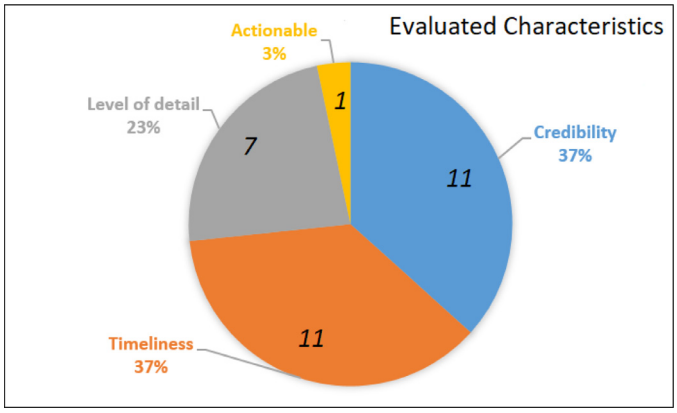

Fig. 3. Importance of the evaluated characteristics.

among the top-5 types of cyber threat intelligence, the demand for intelligence on *vulnerabilities and exploits* as well as *malwares* is generally 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. 5, by applying a threshold-based approach.

Since the types (iii - v) comprise many sources (i.e. user accounts) for which the same metadata (i.e. features) is available, the experts were asked to select the features (which are explained in more detail in the following subsection) 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 4.3). On the basis of such insights, 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 websites were rated higher with regards to the level of detail, Twitter is seen as a much more timely source type. In addition, on Twitter users decide for themselves which content (Tweets) to publish, what their profile's description should be and which other users they want to be directly connected to (following). There are no fixed categories for content or users since they are usually linked to certain topics through the use of hashtags. Combining these findings and the fact that Twitter provides unified metadata on each user, which allows for better assessment and comparison, the authors decided to investigate on Twitter as the OSINT source.

4.3. Feature selection

From the analysis of the related work, resulted that all existing methodologies aimed to identify cyber threat intelligence in different forms and qualities using natural language processing (NLP) and machine learning (ML) techniques. Therefore, no investigations about the features of the sources themselves, that provide CTI-related information, are known to the authors. Only few of them examined aspects of the source but none of them apply to the sources theirselves a feature-driven ML approach. On the other hand, various efforts were conducted to examine the role and characteristics of influencers on Twitter, such as, users who are 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 metadata, the social graph and textual data from tweets.

Based on such information, the first full set of features, which constists of the 26 metadata sources listed in Table 2, has been selected by considering 3 aspects:

- *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 of followers* and so on). Some of these features are transformed into numeric values, in order to have a quantitative computation and

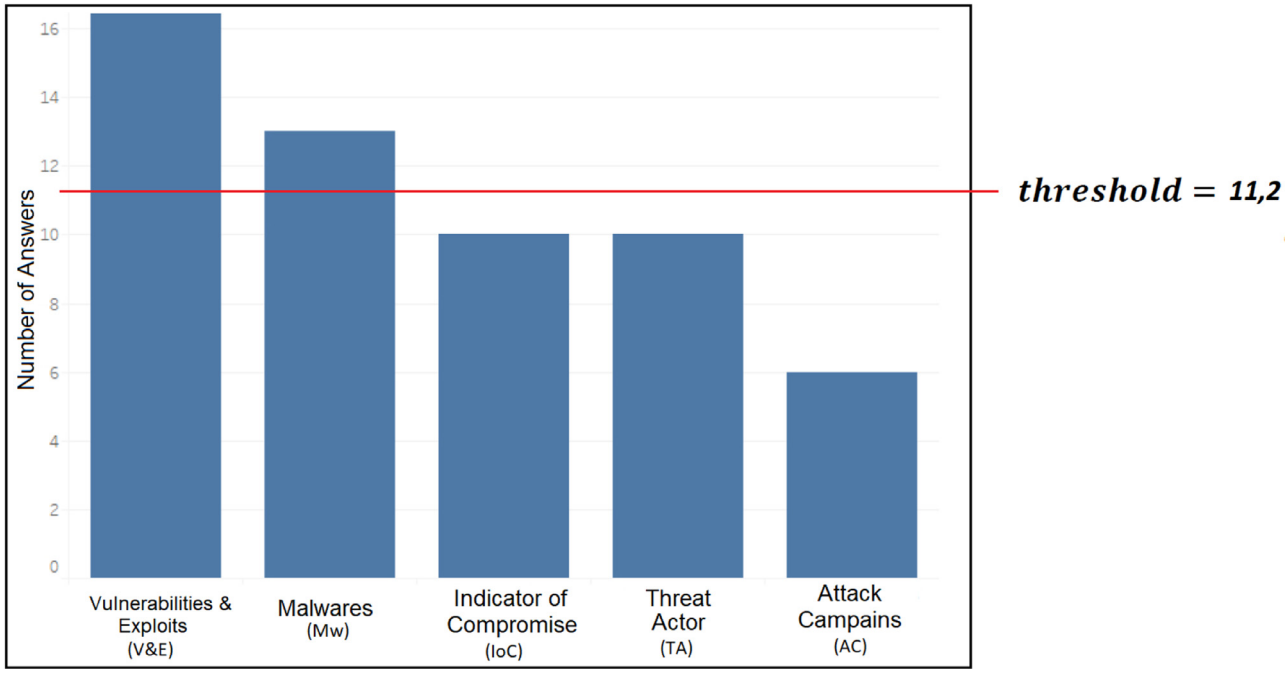

Fig. 4. CTI types - The red threshold represents the mean value, which was obtained by dividing the sum of the "Number of Answers" of all CTI types by the number of "CTI types". (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

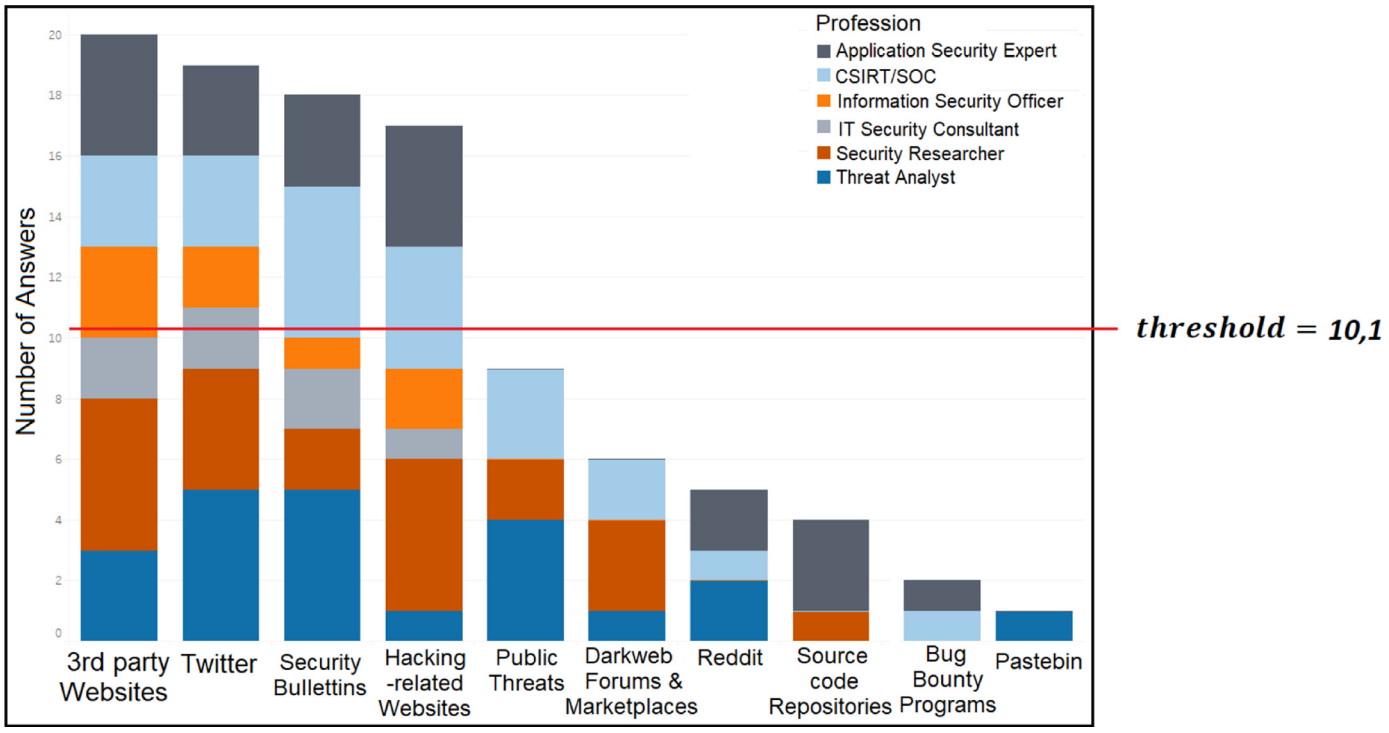

Fig. 5. CTI sources - The red threshold represents the mean value, which has been obtained by dividing the sum of the "Number of Answers" of all CTI sources by the number of "CTI types". (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Selected features based 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

comparison such as the length of the user's profile description (bio), or calculated from the profile metadata, such as the ratio of users following a certain account (followers) to the number of users that are being followed by this account (following).

- *Social graph related features:* they are related to the connections (edges) among certain users (i.e. nodes) and allow to inspect the relations between them within a group or community of connected profiles. Twitter provides three different kinds of social graphs based on: (i) *followed/following*, (ii) *retweets* and (iii)*mentioned/mentions*, where in-degree and out-degree values of each node can be computed as integer values that can then be used as features and compared, as described in Subbian and Melville (2011), Alrubaian et al. (2017), Bakshy et al. (2011). In particular, the graph (i) consists of the connections that occur between 2 profiles if one user is following another user and vice-versa. Analogously to it, the two other graphs describe the relations between users that (ii) retweet each other's tweets or (iii) mention each other using a user's Twitter handle/operator,

i.e. the username is preceded by the '@' sign. The data necessary to build social graphs (ii) and (iii) includes the metadata of the Tweets published by the involved users since this data contains the mentions and retweets.

- *Tweet related features:* other features can also be generated from the metadata of each user's Tweet. This data includes, among others, the time the Tweet was published and the hashtags, URLs and mentions it contains. While it requires the collection of more data, this feature type can provide additional information on the user's behaviour with regards to the published Tweets. For example, the mean time between Tweets during an observed time period can be calculated as well as the average number of hashtags used per Tweet.

As is evident, some of the proposed features are quite intuitive and widely used in other related works (e.g. the *num_followers, num_mentions*), while other features (e.g. *length_bio, length_username, length_url*) have been underestimated in other previous works. The latter constitute in our case further strengths, as the contribute to make our model unique and to be able to achieve high performance, as the results gathered from the conducted evaluation show (see Section 5). That's why, some features related to the length of the content were considered, for example: *length_bio*: the longer it is, the more information it contains, in fact, usually the hackers or those who want to use an account for a short period, they do not put a lot of effort to compile it; *length_username*: the longer it is, the more difficult is to be discovered, that's why some hackers register profile one-way-use, and the short accounts are typically already taken; similarly the of the *length_url*: the longer it is, the more difficult is to be discovered.

Such first feature-set has been engineered only from meta data of the Twitter profile and its Tweets, with the purpose of testing whether metadata-based features are sufficient to assess the relevance of a Twitter profile as a cyber threat intelligence source. The textual content of a user's Tweets is intentionally not used in the first feature-set. It is used, instead, to generate an independent set of content-based features, which in turn requires additional preprocessing steps and natural language processing techniques, as it has been shown in Zhu and Dumitras (2018), Khodabakhsh et al. (2018), Nebot et al. (2018), Jain et al. (2018).

It is a more expensive approach in terms of computational costs, but it has also shown promising results in the related work and it is even suggested for further research in the field of cyber threat intelligence (Le et al., 2019). This means that the second feature set not only is suitable for comparison with the first feature set in this work but it also extends recent research on the identification of relevant CTI.

As suggested in Le et al. (2019), word embeddings are examined as a feature set that represents all the textual content, i.e. Tweets, of a user in a high-dimensional, continuous vector space. This technique strives when determining the similarity between different textual data. Le et al. suggested to use this advantage for the identification of Tweets containing relevant cyber threat intelligence among other, non-relevant Tweets through the use of novelty classifiers. Based on this idea, the word embedding feature set is used to determine the similarity of a potentially relevant CTI source and known, truly relevant CTI source. Such 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 Nebot et al. (2018), that strives when determining the similarity between different textual data.

The configuration details regarding the parameterization of this feature engineering step are reported in Table 3, which have been established according to the related works and through manual tuning.

Table 3
Word Embedding Feature set configuration parameter.

Parameter	Description
dm=1	distributed memory (PV-DM) algorithm
vector_size=50	dimensionality of the feature vector
windows=2	maximum distance between the current and the predicted word in a sentence
min_count=8	all word with lower frequency will be ignored
alpha=0.065	initial learning rate
min_alpha=0.065	prevent the learning to drop any further

4.4. Regression-based scoring

As stated at the beginning of this section, this methodology aims to predict/provide a numeric value quantifying each Twitter user's relevance as a CTI source based on their features. To achieve that, a AI-based approach centered on Supervised Machine Learning techniques has been adopted. This means that, starting from a initial data set, a part of it (i.e. training data) is used to define a classification model on the basis of the features that we choose and proposed on the basis of our review analysis in combination with our intuitions, while another part of data (i.e. test data) is used to evaluate the goodness and performance of such derived model. In our case, the timeliness of each intelligence is used in combination with the total count of CTI published by a source to assign the CTI Relevance Score (by using proposed Eqs. 1 and 2, that are below described) to each source in the training data, on the basis of the above described features. This labeled training data is then used to train regression models for predicting the CTI Relevance Score for each source in the test dataset. The meta-data feature set, that has been described in Section 4.3, consists of both static (e.g. "length_username") and dynamic features (e.g. all information regarding retweets, mentions, etc.) between which no further distinction is made. Since the timeliness of the observed intelligence is essential for the CTI Relevance Score, any changes or additions to the training data can lead to variations in the predicted scores. Of course, as is well-known in the field of machine learning, adding a new source to the training data which published certain intelligence much earlier than all the existing sources in the training data would require the whole data set to be relabeled, i.e. recalculating the CTI Relevance Score, and could therefore change the resulting regression model. But such research aspect, regarding the automatic adaptation of the classification model over the time as described by Kauschke et al. in Kauschke and Fürnkranz (2018), which focuses on the adaption of existing classification models to new data, falls outside the scope of the current research, and consequently it is not considered.

Instead, as already mentioned, such a supervised machine learning approach requires a labeled data set which means that the score needs to be defined first and then calculated for each source in data training data set. A naive score for the relevance could be achieved by using the count of true intelligence that was posted by a source during the observed time period. However, this would ignore two important aspects:

i firstly, the timeliness of threat intelligence is a crucial requirement as it emerged not only from the conducted survey that has been described in Section 4.2 but also from another survey among security professionals from 2018 in Tounsi and Rais (2018). Using only the count of true intelligence would lead to high scores even for sources that mention threats after they are already widely known or even after they are no longer viable threats, e.g. when a vulnerability has been patched for a long time.

ii the second reason why the mere count of true intelligence is not a suitable score is that sources can even be beneficial from

a cyber threat intelligence point of view even if they provide only rare but valid intelligence. Some sources may for example only provide information on threats concerning a certain product or software. Such a source would only provide a low count of intelligence but they would be one of the first to report certain threats and may even be one of very few sources that do so.

Thus, we believe that the score assigned to each threat intelligence source should be based on the weighted count of all true intelligence by this source.

In particular, I represents a threat intelligence source, whereas r_i is a published intelligence published from I. In order to support the evaluation of the relevance of a threat intelligence source, a score function has been 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 represented through Eq. (1).

The time frame $[0; s \cdot C^{1.25}]$ is split into C intervals and scores are computed as a step function resulting in values between [1; 0.5]. The time t affects the value of c which is the number of the time interval that contains t. Let us say the time t would be within the bounds of the second time interval $[s \cdot (2-1)^{1.25}; s \cdot 2^{1.25}]$ then is follows that $c = 2$. Note the range for which c is defined. Its maximum is equal to C such that the top function is defined on the interval $[0; s \cdot C^{1.25})$. The upper bound of this interval is reached for $c = C$. This value is also the lower bound of the interval for the bottom function: $[s \cdot C^{1.25}; \text{inf})$.

$$r_i = score(t_i) = \begin{cases} 1 - 0.5\left(\frac{c}{C-1}\right)^2, c = \left\lfloor \left(\frac{t}{s}\right)^{\frac{1}{1.25}} \right\rfloor & \text{if } t < s \cdot C^{1.25} \\ 0.5^{1+\frac{t-s \cdot C^{1.25}}{s}} & \text{if } t \geq s \cdot C^{1.25} \end{cases} \quad (1)$$

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 $s \cdot C$, the score is score$(s \cdot C) = 0.5$ and intelligence mentioned later than this point of time gets a score below 0.5 assigned through the exponential decay function. In other words, for example, we choose the size of the first interval $s = 86{,}400$ (i.e. one day) and the number of intervals to be $C = 5$. If a CTI source publishes some intelligence, for example 32 hours (t = 115,200) after its very first occurrence (by a different source), it falls into the second interval $[s; s \cdot 2^{1.25}]$ and gets the same score assign as all the other intelligence published within this time interval: $score(t) = 1 - 0.5 \cdot (1/(5-1))^2 = 1 - 0.5 \cdot (1/4)^2 = 0.97$. An additional numerical example, in correspondence with $t = s \cdot C^{1.25}$ (that means when the case of the equation is switching case) is given. As a consequence, by replacing the equation's parameters with the appropriate values, the resulting score is calculated as: $score(s \cdot C^{1.25}) = 0.5^{1+\frac{t-s \cdot C^{1.25}}{s}} = 0.5^{1+\frac{s \cdot C^{1.25} - s \cdot C^{1.25}}{s}} = 0.5^{1+\frac{0}{s}} = 0.5$. 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 Eq. (2).

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

In particular, R represents the full set of all scores and R_I the scores for intelligence shared by source I. The arithmetic mean is calculated over all single relevance scores $r_i = score(t_i)$ of a source $R_I = \{r_1, r_2, \ldots, r_I\}$ and weighted by the logarithmically normalized number of threat-related terms that have been observed for the source I. Then, after that a CTI Relevance Score is assigned to all sources, both sources (characterized by the identified features) along with their CTI Relevance Score are used as input data in order to train a regression model. This model is then used to predict the relevance (see Section 5), which is measured through a value between [0,1], of other sources (i.e. other instances).

In our case, each Twitter user represents a potential threat intelligence source, and we collected all their Tweets from a three-year time period up until the time of prediction as well as the user's meta-data as provided by the Twitter API at said point in time. All sources (i.e. users) were split into training and test data sets using cross-validation (see Section 5.3) and the training data set was labeled by assigning every source its CTI Relevance Score calculated from the single relevance scores of each Tweet containing threat intelligence. A regression model is then trained to predict the CTI Relevance Score for each source in the test data based on either word embedding or meta-data features. Since the time of prediction coincides with the end of the data collection period, every collected Tweet and all meta-data present at that point in time can be included in the training process. The single relevance scores of a Tweet (r_i) are not determined when the Tweet was seen, but when the CTI Relevance Score is calculated to label the training data, i.e., after all data has been collected. The single scores are also neither used as a feature for the training nor are they predicted by the model. Thus, no temporal bias as described by Pendlebury et al. in Pendlebury et al. (2019) is to be expected based on the collected data since training and evaluation of the regression model happens at the same point in time and predictions are not simulated to have happened before the end of the collection period. However, it should be noted that further collection of future data would of course change the CTI Relevance Scores calculated at some point in the future.

Finally, the trained model is used to predict score for cyber threat intelligence sources in the test data set. The result of such prediction is then employed to take decision when generating alerts or selecting new sources to be monitored.

5. Implementation and conducted experiments

In this Section, the data collection approach is first presented, second the used regressor models and the adopted evaluation metrics are described and then the obtained experimental results are reported and discussed.

5.1. Data collection

The data collection is focused on Tweets and Twitter profiles, including metadata, related to the field of cybersecurity as a starting point for generating sets of training and testing data later on. As result of the survey, an *initial list* of cyber security and cyber threat-related hashtags, consisting of the following terms *infosec, cybersecurity, security, threatintel, hacking, malware*, was manually compiled. 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, assumed

to be relevant to the field of cyber threat intelligence. This procedure was repeated on a daily basis from the 1st until the 31st of May 2019, and the resulting extended list is reported in the box below reported.

> Extended list: *bugbounty, cve, cvss, cyberattack, cybercrime, cybercriminals, cybersec, databreach, dataleak, exploit, exploits, hacker, hackers, itsec, itsecurity, privacy, ransomware, redteam, threatintelligence, virus, vuln, vulnerabilities, vulnerability.*

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 the removal of the duplicates, 156 Twitter profiles remained that were added with further 16 Twitter profiles used in Le et al. (2019), by reaching a total of 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 that, these users were related to the domain of technology and used similar vocabulary but they did not focus 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 Twitter and stored in JSON files. Those files were imported in a Data-as-a-Service platform called "Dremio" to be queried using the Structured Query Language (SQL).

After that, the followng pre-processing steps have been applied in order to obtain the potentially threat-related terms in all Tweets: (1) Removal of emojis and other non-ascii characters since the data is limited to English Tweets. (2) Tokenization of the Tweet using the tokenizer trained on the NLTK (Natural Language Toolkit) Twitter corpus. (3) Filtering of all tokens which are 'mentions', i.e. Twitter usernames or URLs that are already detected by Twitter and listed in the Tweet's meta data. (4) Removal of all English stopwords and punctuation including the pound sign (#) from hashtags. (5) Removal of genitive endings ('s) in all tokens. (6) Converting all tokens to lower case and applying the lemmatisation technique of NLTK. (7) Check for threat-related keywords in the resulting tokens. If no keyword is found, the Tweet is assumed to be not related to threat intelligence and dropped. (8) Split compound words which consist only of common and domain-specific words already found from the data set of the previous year. For example 'DNS-based' is split into 'DNS' and 'based'. (9) Filtering of all common English words and domain-specific terms. (10) Filtering of all threat-related keywords such that only unknown terms remain.

As described in the comparable system 'DISCOVER' (Sapienza et al., 2018) only Tweets which contain threat-related keywords are considered for term mining, otherwise the Tweet is dropped at the seventh step. Since the keywords used in 'DISCOVER' are not listed in such work, new threat-related keywords are manually compiled from two third-party websites. All words listed in the glossary of "cybrary.it" Cybrary.it are parsed from the website and merged with all keywords extracted from the Common Attack Pattern Enumeration and Classifications (CAPEC) by the non-for-profit organization MITRE Corporation MITRE. Therefore, all common words from the English corpus of the NLTK are removed from the CAPEC catalog and all remaining terms are considered keywords. All terms found by the term miner which appeared at least twice are then stored in a JSON file along with the date and time of the term's first occurrence for each Twitter profile. The potential CTI terms need to be verified using external information from Open Threat Exchange (OTX) as introduced in Section 4.1. Information on malware families, threat actors and similar cyber threat intelligence is provided by OTX as Structured Threat Information Expression (STIX) OASIS Open and can be parsed in Python with the official module 'cti-python-stix2' (version 1.1.2). The OTX platform provides an official Python Software Development Kit (SDK) called 'OTX-Python-SDK' to query the API. For each term a single request is made to search OTX objects that contain the term in the name, description, tags or adversary field. If such an object is returned by the API the term is considered a valid cyber threat intelligence.

5.2. Regression models and evaluation criteria

This Section is related to the last part of the methodology, where both feature sets are used to train multiple regression models for predicting the CTI Relevance Score of a CTI source, in this case Twitter profiles. The goal is to find the best regression model for this prediction and use it to improve the time advantage of early alerts from sources with a high score, as it is described in Section 4. To ensure comparability across both feature sets, all the experimented regression models have been trained separately on each feature set, and then they have been evaluated using the same metrics. This section explains the selection of the regression algorithms in comparison to the related works, where no regression approach was used in the context of CTI and for Twitter as an OSINT source.

In our case, 5 regression algorithms were evaluated and compared to the related works. In particular, the following ones have been chosen as the no-regression version is typically applied in the related works: (i) SVM Regressor (SVR) by applying the Gaussian Radial Basis Function (RBF) kernel; (ii) Random Forest Regressor (RFR); (iii) a Gradient Boosting Tree regression (GBTR) model, (iv) the Extra Trees Regressor(ETR); and (v) a Multi-Layer Perceptron regressor (MLPR).

Specifically, Support Vector Machines (SVMs) and Random Forest (RF) classifiers have been used in related approaches (Le et al., 2019; Khodabakhsh et al., 2018; Nebot et al., 2018) as supervised machine learning techniques, because from one side, a SVM can typically handle higher dimensional data due to the kernel trick, whereas the Random Forest algorithm, which is based on decision tree ensembles, can reduce overfitting by using cross-validation (Sabottke et al., 2015). As a consequence, two related regression algorithms that is a SVM Regressor (SVR) by applying the Gaussian Radial Basis Function (RBF) kernel and Random Forest Regressor (RFR) have been used to establish a baseline for comparison. Then a Gradient Boosting Tree (GBT) regression model has been selected, as the GBT algorithm showed well performances to classify real-time events on Twitter by using heterogeneous features (Khodabakhsh et al., 2018; Ke et al., 2017), as well as the Extra Trees Regressor(ETR), which is less susceptible to overfitting (Geurts et al., 2006), that extends the Random Forest Regressor has been choosen. Furthermore, as two feedforward artificial neural networks on word embeddings derived from Tweets, namely the Convolutional Neural Network (CNN) and the Multi-Layer Perceptron (MLP), where evaluated in Nebot et al. (2018), with only

Table 4
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 Nebot et al. (2018).
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 Devore (2012).
GBTR	500 boosting stages during training optimizing the least squares loss function and limiting the maximum depth to 5 nodes per tree, as in Ke et al. (2017).
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 Nebot et al. (2018).

a slightly different (0.05%) regarding the mean average precision (MAP) and it was also evaluated in Le et al. (2019), only a Multi-Layer Perceptron regressor (MLPR) has been considered. While the term frequency-inverse document frequency (TF-IDF) features is used in Le et al. (2019), in this current work word embeddings for novelty classification approach is adopted to calculate a relevance score for CTI sources based on the distance or similarity of a new source to the CTI community's centroid. This is because the Tweets containing true CTI are assumed to form a cluster which is separable from other Tweets as argued in Lee et al. (2017). Both feature sets, the source meta-data features and the word embedding features, proposed in the previous section are used. The implementation of the regression models was based on "scikit-learn" Python library Scikit-learn, and the configuration paramters are reported in Table 4.

The implemented evaluation criteria were based on the following metrics, which have been used to evaluate the performance of regression models:

- *Mean Squared Error (MSE)*: it is typically used to express the performance of a regression model by computing the arithmetic mean of all squared errors that have been made during the prediction of a numeric value;
- *Coefficient of Determination (R^2)*: it is used to assess how well a regression model fits the data set, and it represents the proportion of the variance in the dependent variable that is predictable from the features that the model was trained on Marsland (2015). The optimal value is $R^2 = 1$ and indicates a perfect fit of the model.
- *Gini Impurity:* ranking the features of a feature set according to their importance for the machine learning task at hand is useful to investigate subsets that can perform similar but are less susceptible to overfitting. The Gini Impurity is a metric to quantify the importance of a feature. It aims to reduce the impurity in each node of a decision tree meaning that all data points represented by that node should belong to the same class and no data points are mismatched (Marsland, 2015).

In the following subsection, the results are presented and discussed.

5.3. Results discussion

This section presents the evaluation of the methodology proposed in Section 4 and a comparison of the approach with the related work to determine the potential improvement on the timeliness of alerts generated by existing CTI systems. As already discussed in Section 2, only very few metrics for the assessment of CTI sources have already been presented in recent research and they are so far only used to select the top-ranking sources for further monitoring and CTI collection. However, they were calculated on the data set in this work for comparison to the CTI Relevance Score. As a baseline, the unweighted count of true intelligence per source is used, as already discussed in Section 4. The Utility Score is introduced in Sabottke et al. (2015) and favors sources that provide intelligence about exploitable vulnerabilities with an assigned CVE identifier. In Schaberreiter et al. (2019), Schaberreiter et al. introduced a methodology to evaluate the trust in the quality of CTI, which focuses on various aspects of structured CTI, e.g. STIX or OpenIOC, and is not directly applicable to the detection of emerging CTI terms from unstructured data source like Twitter. However, they described a timeliness parameter as part of their metric that is compared to the CTI Relevance Score. The experiments exploited the list of 659 CTI terms that were found across all 230 selected sources (i.e. Twitter accounts).

The graph represented in Fig. 6 shows the different scores for the labeled training dataset used in this work, where every instance of CTI has been verified and labeled based on information from Open Threat Exchange (OTX). In particular, it visualizes the [0, 1]-normalized scores for all sources that are sorted on the horizontal axis according to their true intelligence count (baseline). The baseline (red) shows that the left-most sources are CTI sources that had high counts of true intelligence while other Twitter accounts like 'opera' shared no or only very few cyber threat intelligence. Comparing the baseline with the timeliness parameter (green) shows that some sources have published the same number of CTI indicated by a small plateau in the red graph but differ with regards to the timeliness of their intelligence, i.e. noticeable differences in the green graph for the same sources. The CTI Relevance Score (blue) can be considered a combination of the intelligence count and the timeliness parameter with the difference that it applies a non-linear decay function to the timeliness of intelligence as opposed to the timeliness parameter. It focuses more heavily on the timeliness which decreases more rapidly during the initial time period than with a linear approach and can be observed as higher variations in the scores. In the right segment of the graph (close to 'airbuscyber'), a noteworthy peak of the CTI Relevance Score can be observed, which shows that even sources which rarely publish intelligence (i.e. have a low overall count) still received a notably higher CTI Relevance score (blue line) compared to the other scoring functions as long as the intelligence was published in a timely manner (i.e. among the first sources to report it). In other words, the focus on timeliness is once more highlighted. In this case, it can be seen that, even though two sources have a low count of true intelligence (red graph), they were among the first or maybe the only sources within the community that shared certain intelligence. Such behavior is especially relevant for sources that focus on CTI for niche software etc. That is to say, such sources might only share CTI for a certain product or regarding a certain threat that can still be of interest for security experts. Using the other scores the source could easily be overlooked. As already mentioned above the Utility Score (orange) ignores the timeliness of intelligence but focuses on intelligence regarding vulnerabilities (CVE) with high exploitability. Examining the peak in the graph for the source 'inj3ct0r' suggests that this source provides high counts of intelligence containing CVE identifiers for exploitable vulnerabilities. It fits in the picture that this source has

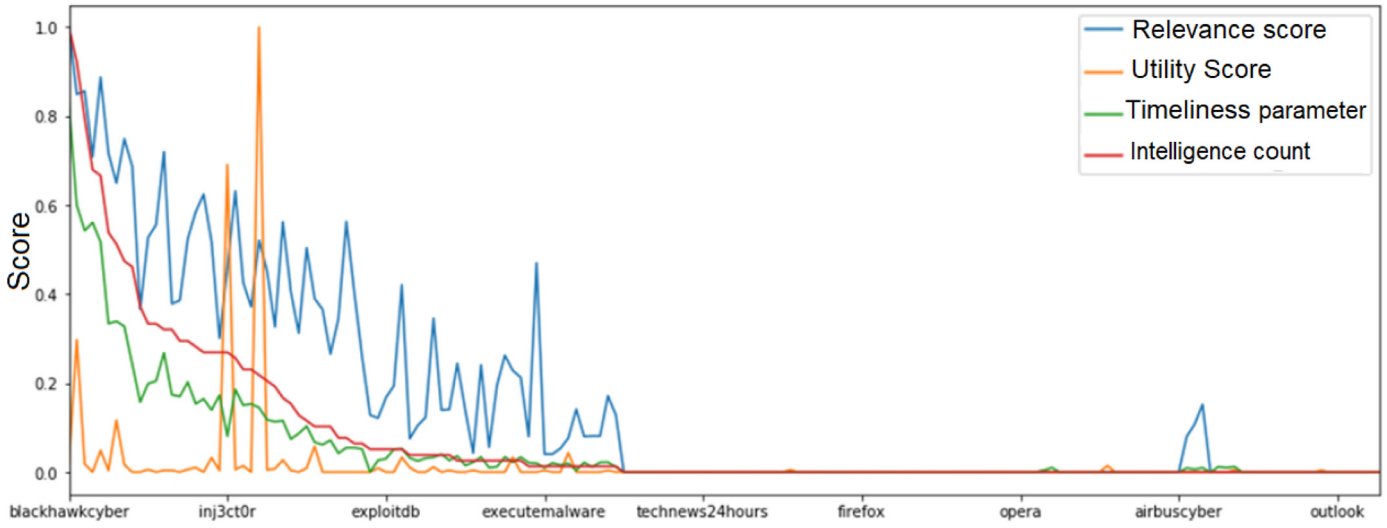

Fig. 6. CTI-Relevance-Score.

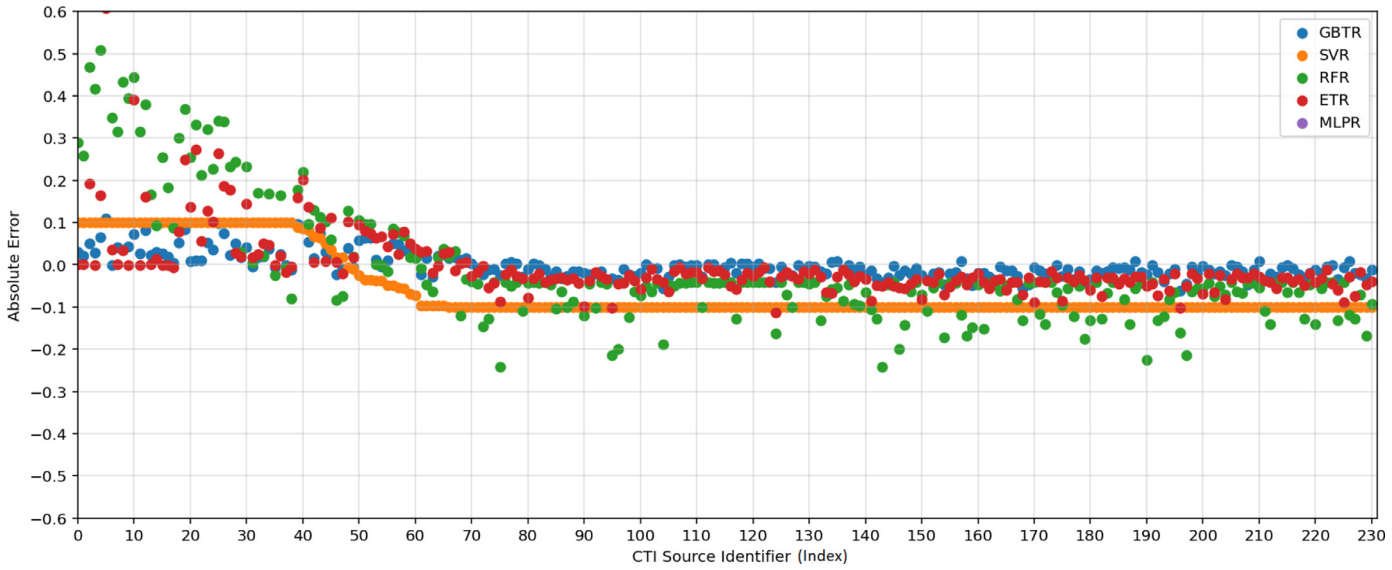

Fig. 7. Absolute Error.

a lower timeliness compared to its neighbors in the graph since it is to be expected that the investigation of the exploitability of CVE takes a certain amount of time and delays the sharing of such intelligence.

Overall, the CTI Relevance Score seems to be comparable to the timeliness parameter but favors very fast CTI sources more heavily and can therefore detect sources that provide rare CTI for niche products or only regarding the source's domain of expertise.

Whereas, Fig. 7 shows the Absolute Error between the real value and the predicted ones, by sorting the sources (x-axis) in descending order based on their intelligence count. It is worth noting that, after 190 sources the remainder of sources had no intelligence published at all and the errors in the predictions did not differ notably from the previous ones. 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 Sabottke et al. (2015); Zhu and Dumitras (2018), meaning that all data was separated into ten partitions of equal size and each partition was used for evaluation once while the different models were trained on the remaining data. The full source meta data feature set was examined first to establish a baseline before looking into feature subsets and the word embedding feature set as an alternative. All models predict a normalized value between zero and one as the CTI Relevance Score with high scores indicating high relevance as a cyber threat intelligence source. The absolute errors showed in Fig. 7 depicts the deviation between the predicted score and the true score for each instance of the data set, i.e. each Twitter user. Since the predicted score can be above or below the true score, the error is negative if the prediction is less than the true value. Data points close to the horizontal centerline representing zero deviation are more accurate predictions. All sources are numbered as shown on the "Source Index" axis and are also sorted by their true CTI Relevance Score such that the source with the best score has index zero and the score decreases along the axis until only sources with a score of zero remain from index 50 onwards. It can be observed that the Gradient Boosting Tree Regressor (blue) performs best with a maximum error $err_{\max} = 0.12$ while the other two decision tree based models, Random Forest Regressor (green) and Extra Trees Regressor (red), result in large errors for instances with a high score as can be seen on the left end of the figure. This makes them less

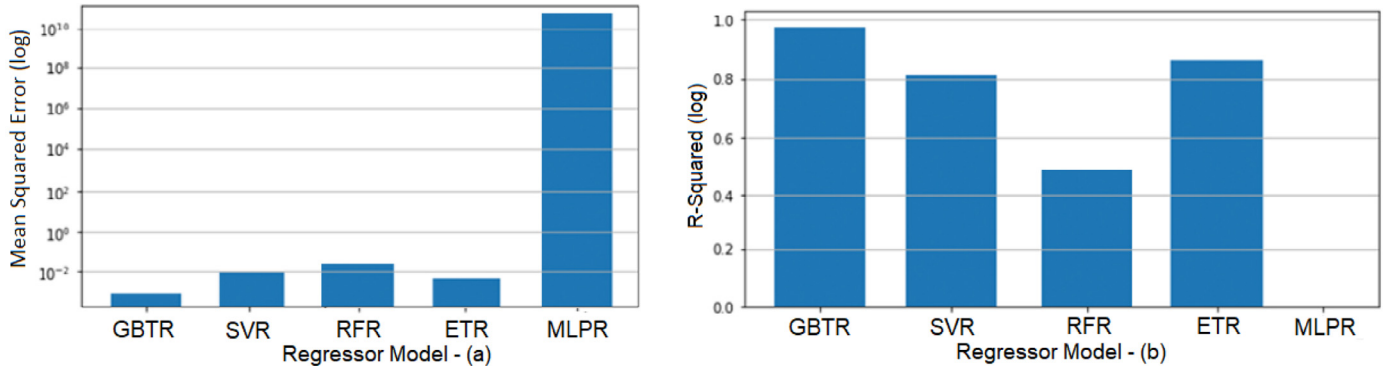

Fig. 8. Mean Squared Error (MSE) for the score prediction of each regression model on the source meta data feature set AND R^2 for the score prediction on the meta data feature set.

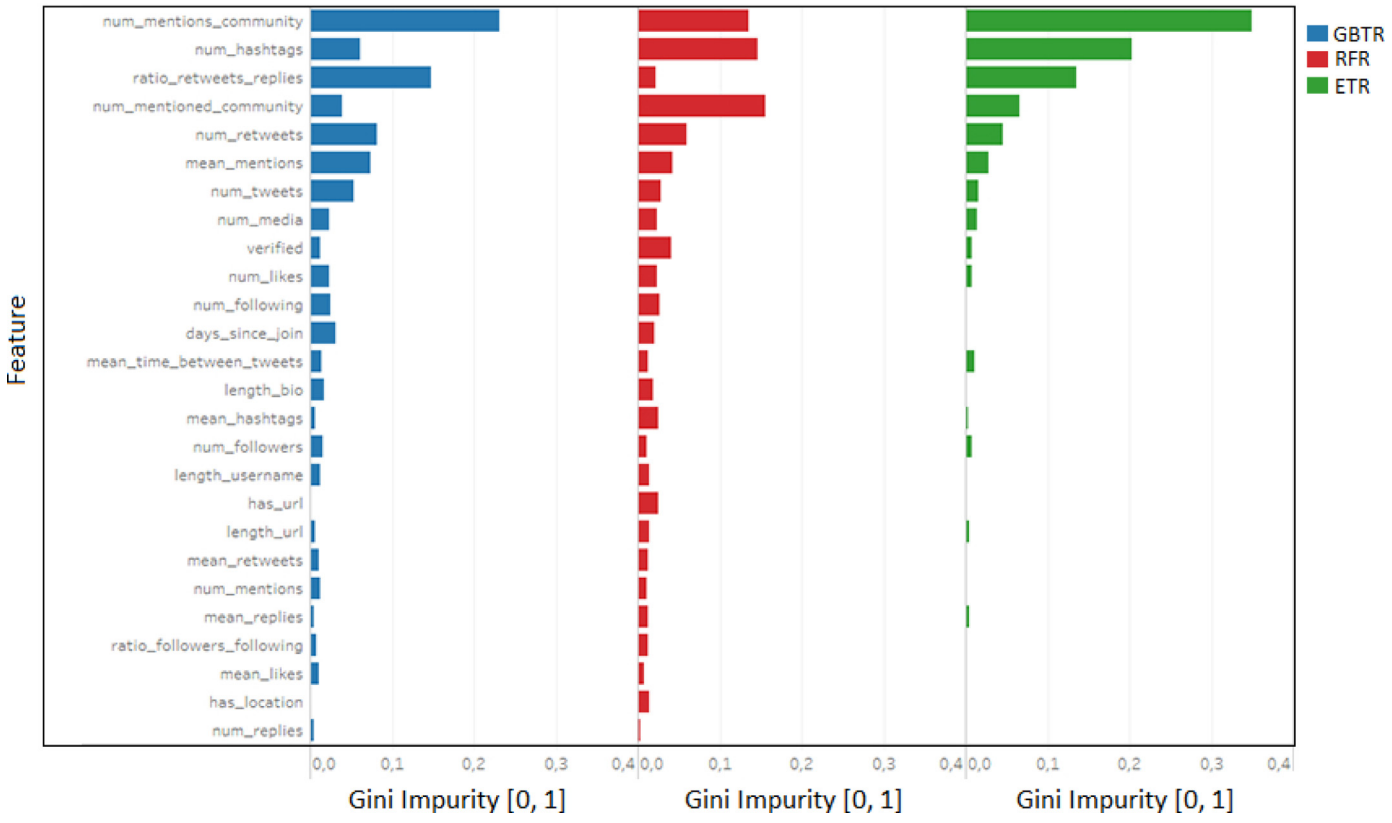

Fig. 9. A comparison of the feature importance for each regression model (i.e. GBTR, RFR and ETR) on the source metadata feature set based on the entropy level by using the Gini Index criterion.

favourable for the task since one would be primarily interested in the high scorers as potential threat intelligence sources. 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. This is also shown on the log-normalized axis for the MSE in Fig. 8-(a).

How well a regression model fits the true data points is typically described by the Coefficient of Determination (R^2). 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 Fig. 8-(b). One characteristic of R^2 is that its value increases monotonously when adding features to the model (Devore, 2012). There are many ways to reduce the risk of overfitting when training a machine learning model. However, the authors focused on comparability with the related work and implemented the measures mentioned in these works as follows. The training and test data was selected using 10-fold cross-validation to avoid overfitting and in accordance with the related work. The number of features for the word embedding feature set was selected analogous to the approaches in the related work which also operated on Twitter data sets described in Section 4.3. For the meta-data feature set potential overfitting was reduced through the selection of features based on their feature importance as described in the following paragraph.

For the selection of such feature subsets all features are ranked according to their Gini Impurity for the decision tree based models GBTR, RFR and ETR. These results are displayed in Fig. 9 sorted by their Gini Impurity averaged over the three models. For each of the three models the top-ranking f_{num} features were combined into feature subsets which were then used to train new models. The value for f_{num} was determined via the maximization of

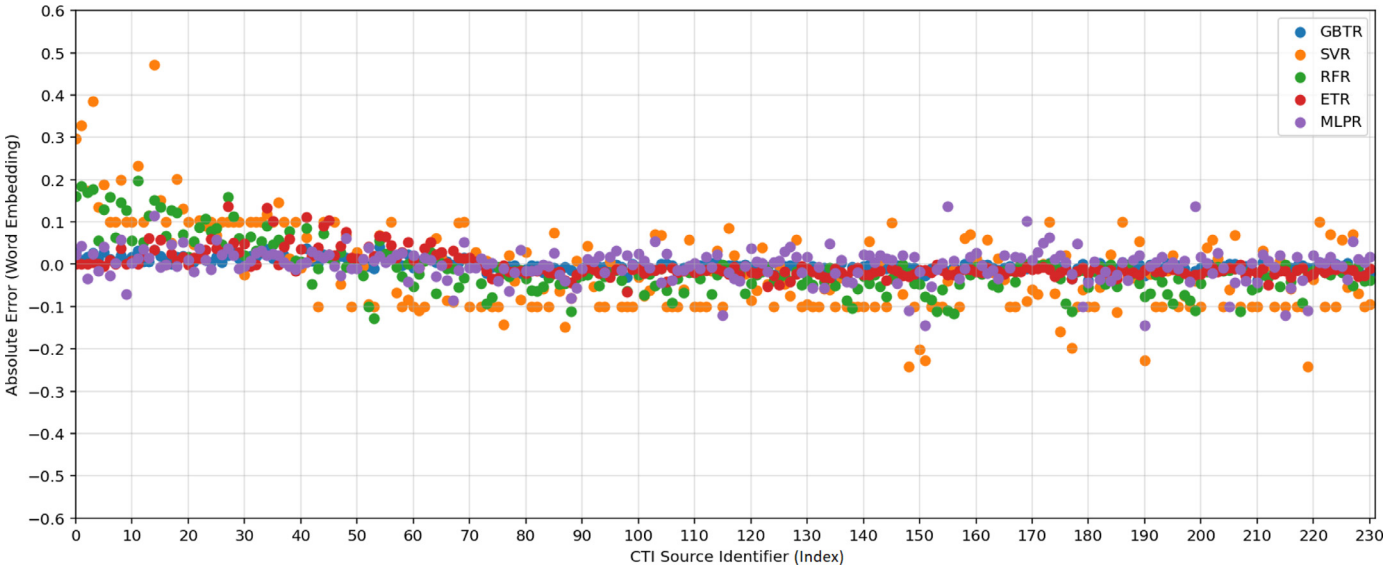

Fig. 10. Absolute error for the score prediction of each regression model on the word embedding feature set.

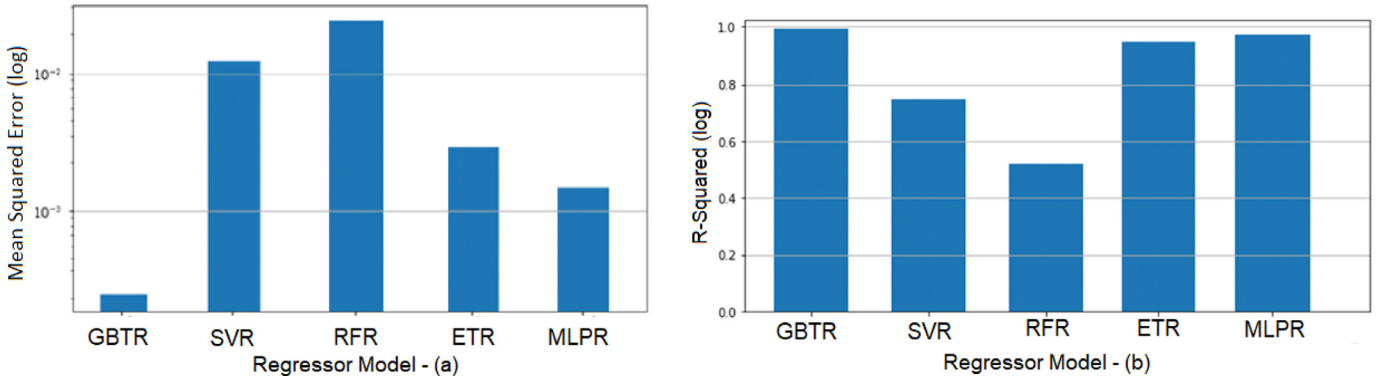

Fig. 11. Mean Squared Error and R^2 for the score prediction on Word Embedding feature set.

R^2 for each new model. The best value was R^2_{max}= 0.77 which was achieved with f_{num}=8, i.e. the top eight most important features, from the Extra Trees Regression model using the Gradient Boosting Tree Regressor and resulted in the following feature subset: *num_mentioned_community, num_mentions_community, num_retweets, ratio_retweets_replies, num_media, mean_mentions, mean_hashtags, mean_time_between_tweets*.

A closer examination of the feature subset shows that the features based on the in-degree (*num_mentioned_community*) and out-degree (*num_mentions_community*) of the source's social graph are of high importance. All in all the meta data of the source's Tweets seems to be more influential for the trained model compared to static meta data of the source like the location, url or whether the account has the "verified" status. Interestingly, the number of Tweets containing media, e.g. pictures, was ranked rather high for this model. In the end the presented feature subset was able to achieve a similar performance as the full feature set with a slight decrease of 21% in R^2 for the Gradient Boosting Tree Regression model by applying the Occam's Razor principal removing additional nodes, i.e. features, from the decision tree-based models and thereby decreasing the risk of overfitting.

As described in Section 4, all Tweets collected as part of this work's data set were used to train a word embedding model that provides a single feature vector per Twitter source derived from all Tweets this source has posted during the observed time period. This approach was already used for binary classification by Nebot et al. (2018) but for a different topical domain. They achieved good results using word embeddings with dimension 50 for Tweets and in this work the model was trained with the same feature vector size and parameters.

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.

Fig. 10 shows the Absolute Errors when predicting the CTI Relevance Score using word embeddings. Also in this case, only the first 190 sources had intelligence published. However, the errors for the remainder sources did not differ notably from those with low counts of intelligence. With a Mean Average Error of MAE = 0.02 the Gradient Boosting Tree Regressor (blue) again outperformed the other models on this feature set. However, the Multi-Layer Perceptron (purple) shows similar performance. This is in contrast to the results for the source meta data feature set and becomes even clearer when examining the Mean Squared Error (MSE) in Fig. 11-(a) where the MLP regression model improved drastically and ranked second best on the word embedding features.

Fig. 11 -(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 fea-

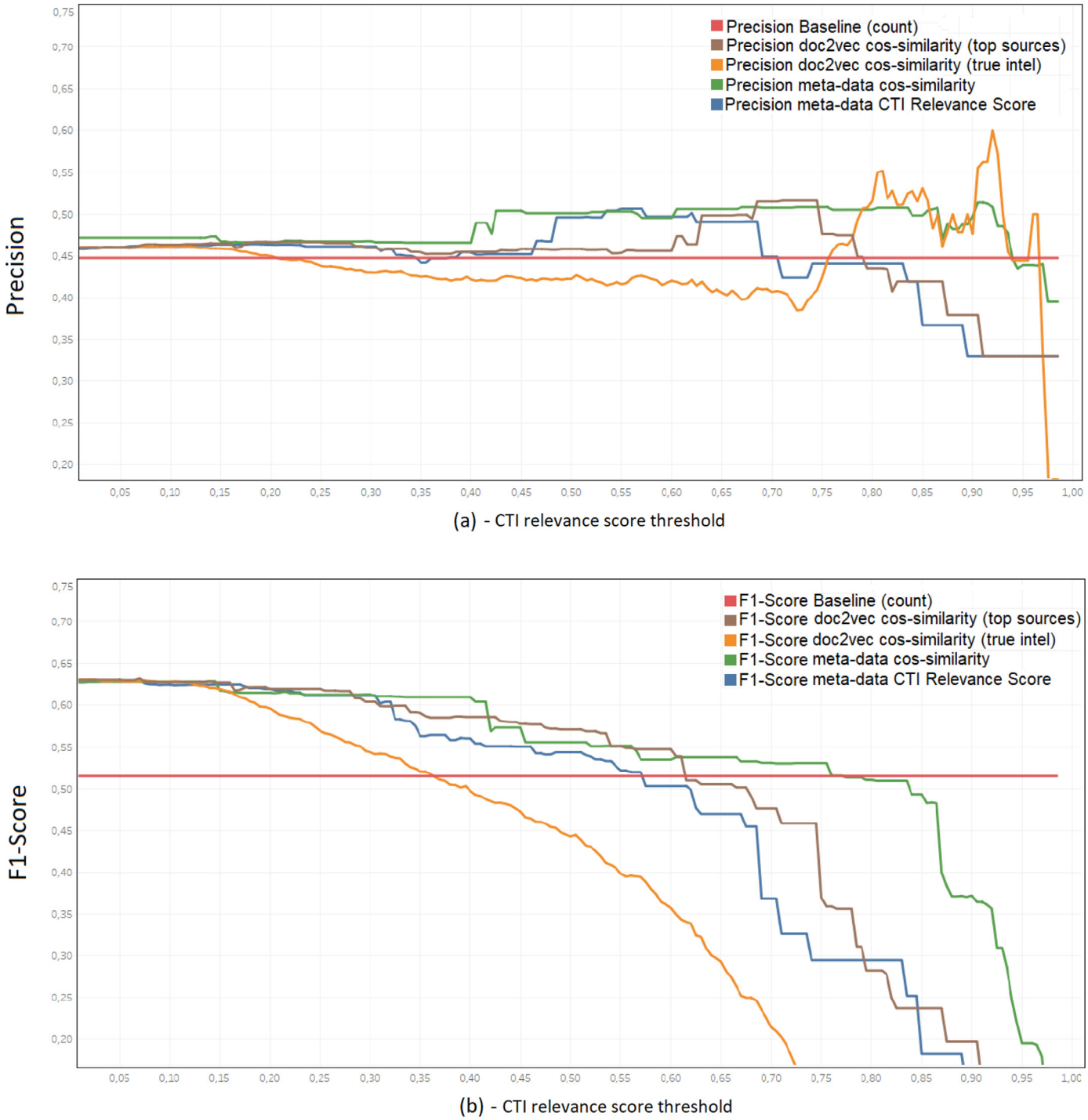

Fig. 12. Precision and F1-Score used to quantify the relevance of the predicted scores.

tures. 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 stage, while maintaining the precision given for a rule set based only on the intelligence itself by ignoring the source. Similar to the method described in Le et al. (2019), since each source can be represented by features derived from its metadata or a word embedding vector, both types of feature sets 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 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, in order to improve the identification of CTI Tweets.

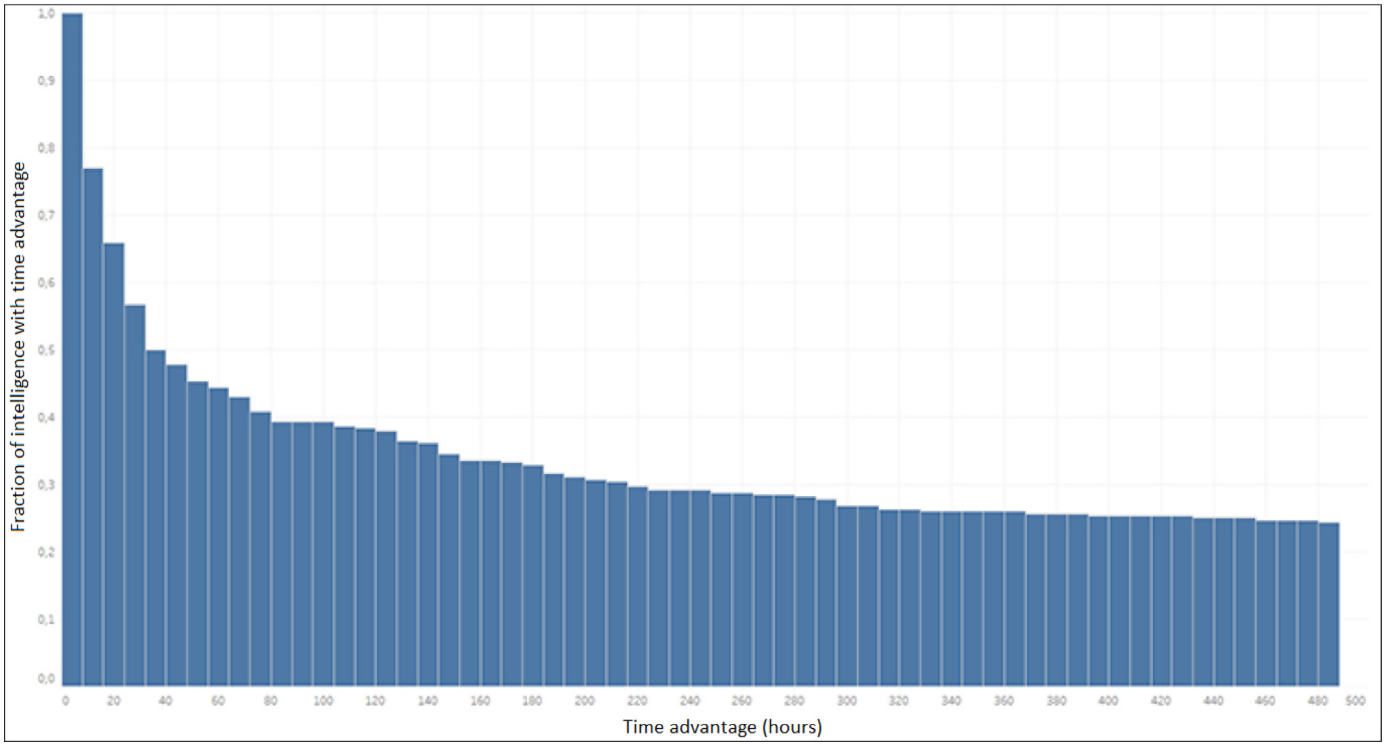

Fig. 13. The time advantage in hours gained when using the relevance score.

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. Since the CTI Relevance Score and the cosine-similarity are both continuous values between zero and one, for both of them a threshold needs to be met by the CTI publishing source in order to be considered relevant and generate an early alert. 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. 12-(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 (Sapienza et al., 2018) 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 true threat 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. 12-(b).

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 for the first centroid (green) has a F1-Score above the baseline from DISCOVER (Sapienza et al., 2018) for all thresholds up to $t = 0.752$.

Through the 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 metadata 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 it from a different source. This time delay in hours is calculated for each instance in the dataset, and showed in Fig. 13, by considering 50% of all intelligence available in the considered dataset, thus achieving a time advantage in comparison to the related work (Sapienza et al., 2018). It shows not only the number of alerts that could be issued earlier, corresponding to their improved timeliness but also the average time advantage to be gained, indeed half of all alerts could have been issued at least 32 hours earlier.

A further observation is that, besides the time advantage, the use of the source relevance can also increase the recall during alert generation. In case certain intelligence regarding niche products or threats is only shared by a single source within the community no alert would be generated at all since none of the other monitored sources verifies the intelligence. To provide an insight into the real world applicability of this work's approach can also be compared to the case study conducted in Sapienza et al. (2018) where the 'DISCOVER' system automatically generated a warning for the emerging "wannacry" ransomware threat. The approach by Sapienza et al., as described in Section 2, monitors the occurrences of threat-related terms and issues a warning based on the constraint that such a term has to be observed multiple times in order to prevent a high false positive rate. For the term 'wannacry' the first warning was generated on the 18th of April 2017 when said constraint was met. However, the term was already observed before this date and could have generated an earlier warning if a different constraint was used to quantify the relevance of this intelligence. We preliminarily investigated the usage of this work's approach to predict the CTI Relevance Score of the source as an alternative constraint. Using the Twitter API the five earliest mentions of "wannacry" as a ransomware were selected and the metadata of the according Twitter user profiles were retrieved. The source metadata features were derived and the Gradient Boosting Trees regression model was applied to predict the $cti\_relevance\_score = 0.466$ for this source. This score lies within the top 12% of all sources in the training data set and in the top 30% of all cyber threat intelligence sources. Such a value indicates at first glance that the source shows characteristics of a real cyber threat intelligence source and the observed threat-related term might likely be relevant intelligence. However, regarding this last observation,

further empirical analysis is needed to investigate the potential influence on the false positive rate in real world systems for cyber threat intelligence.

6. Conclusion

This paper focused on the automated assessment of open-source cyber threat intelligence sources with regards to their relevance as a cyber threat-related source.

A method for the assessment of OSINT sources, driven by features, has been proposed. In particular, (i) a study conducted among researchers and security professionals working in the field of cyber threat intelligence has been conducted in order to highlight the demand for mid-level intelligence on vulnerabilities, exploits and malwares. Furthermore, such study showed that Twitter is seen as one of the most promising OSINT platform with regards to the timeliness of cyber threat intelligence; (ii) a method to collect cyber security-related data set via the Twitter API was realized and the labeling of true intelligence was automated using the Open Threat Exchange as an external data source for verification; (iii) two different feature sets were engineered from the acquired data set, whereas (iv) to quantify the relevance of threat intelligence sources the CTI Relevance Score was formalized and compared to related scores from existing research effort.

Different regression models have been employed for the evaluation of the proposal. From the experiments emerged that the relevance of an open source on Twitter for cyber threat intelligence could be predicted through an automated feature-driven assessment of the source and utilized in different ways.

In particular, the time advantage of early cyber threat detection can be increased when using the quantified source relevance as a decisive factor for automated alert generation in existing systems. Especially for early threat intelligence which often lacks contextual information to be verified it can be beneficial to take the publishing source into account and consider the intelligence relevant if the source has a sufficiently high relevance score itself. This can reduce or even eliminate the need to wait for additional confirmation by other sources in many cases. During evaluation of the threat intelligence found in two years worth of Tweets from 172 cyber security-related users resulted that half of all alerts could have been issued at least 32 hours earlier without impairing the precision, 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.

Besides an improved timeliness of threat intelligence, the source relevance can also lead to an increase in recall of the alert generation which means that more valid alerts (true positives) can be generated from the collected data. Existing systems (Sapienza et al., 2018; Sapienza et al., 2017a; Mittal et al., 2016a; Mittal et al., 2019b) wait until potential intelligence has been observed from multiple sources. This, however, prevents the generation of alerts for highly specific threats that are only published by very few or even just one of the monitored sources. Thus, the score can also be used to select new sources that are relevant, that is, those that have a score above a certain threshold, and which are thereby likely to improve the system's CTI collection.

Several future research works could be conducted on the basis of the current model proposed, for example (i) by adapting and applying the current model to other CTI Sources, thus studying its behavior and associated performance in comparison to current "Twitter" CTI source, as well as (ii) conducting sensitivity analysis in order to understand how "sensitive" such model is when varying the initial data (i.e. dataset) and how predictions change depending on different features.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Andrea Tundis: Supervision, Writing – original draft, Methodology, Validation, Formal analysis. **Samuel Ruppert:** Software, Validation, Formal analysis, Methodology. **Max Mühlhäuser:** Supervision, Validation, Methodology, Formal analysis.

Acknowledgment

This research work was funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE, as well as by the LOEWE initiative (Hesse, Germany) within the emergenCITY centre.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cose.2021.102576.

References

Alrubaian, M., Al-Qurishi, M., Al-Rakhami, M., Hassan, M.M., Alamri, A., 2017. Reputation-based credibility analysis of twitter social network users: reputation-based credibility analysis of twitter social network users 29 (7), e3873. doi:10.1002/cpe.3873.

Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J., 2011. Everyone's an influencer: quantifying influence on twitter. In: Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11. ACM Press, p. 65. doi:10.1145/1935826.1935845.

Berghel, H., 2014. Robert david steele on OSINT. Computer (Long Beach Calif) 47 (7), 76–81. doi:10.1109/MC.2014.191.

Bouwman, X., Griffioen, H., Egbers, J., Doerr, C., Klievink, B., van Eeten, M., 2020. A different cup of TI? the added value of commercial threat intelligence. In: 29th USENIX Security Symposium (USENIX Security 20). USENIX Association, pp. 433–450.

Bridges, R.A., Huffer, K.M.T., Jones, C.L., Iannacone, M.D., Goodall, J.R., 2017. Cybersecurity automated information extraction techniques: Drawbacks of current methods, and enhanced extractors. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, pp. 437–442. doi:10.1109/ICMLA.2017.0-122.

Chen, H., Liu, R., Park, N., Subrahmanian, V.S., 2019. Using twitter to predict when vulnerabilities will be exploited. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, pp. 3143–3152. doi:10.1145/3292500.3330742. KDD '19

Chu, S.C., Kim, Y., 2011. Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites 30 (1), 47–75. doi:10.2501/IJA-30-1-047-075.

Cybrary.it. Cyber security glossary and vocabulary. URL https://www.cybrary.it/glossary/.

Dalziel H., Olson E., Carnall J.. How to define and build an effective cyber threat intelligence capability. Syngress is an imprint of Elsevier, OCLC: 910537102, URL http://www.books24x7.com/marc.asp?bookid=78688.

Devore, J.L., 2012. Probability and statistics for engineering and the sciences, Eighth edition Brooks/Cole, Cengage Learning.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees 63 (1), 3–42. doi:10.1007/s10994-006-6226-1.

Hassan, D., 2018. A text mining approach for evaluating event credibility on twitter. In: 2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE). IEEE, pp. 171–174. doi:10.1109/WETICE.2018.00039.

Jain D., Kustikova M., Darbari M., Gupta R., Mayhew S., et al. Simple features for strong performance on named entity recognition in code-switched twitter data. In: Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching. Association for Computational Linguistics. p. 103–109. 10.18653/v1/W18-3213

Kauschke, S., Fürnkranz, J., 2018. Batchwise patching of classifiers. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, LA, USA, 2–7 February, pp. 3374–3381.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 3146–3154.
Khodabakhsh, M., Kahani, M., Bagheri, E., Noorian, Z., 2018. Detecting life events from twitter based on temporal semantic features. Knowledge-Based Systems 148, 1–16. doi:10.1016/j.knosys.2018.02.021.
Krzysztof, L., Jacek, S.W., Michal, J.L., Amit, G., 2015. AUTOMATED CREDIBILITY ASSESSMENT ON TWITTER. Computer Science 16 (2), 157. doi:10.7494/csci.2015.16.2.157. URL http://journals.agh.edu.pl/csci/article/view/1340
Lahuerta-Otero, E., Cordero-Gutiérrez, R., 2016. Looking for the perfect tweet. the use of data mining techniques to find influencers on twitter. Comput Human Behav 64, 575–583. doi:10.1016/j.chb.2016.07.035. URL https://linkinghub.elsevier.com/retrieve/pii/S0747563216305258
Le, B.D., Wang, G., Nasim, M., Babar, A.. Gathering cyber threat intelligence from twitter using novelty classification.
Lee, K.C., Hsieh, C.H., Wei, L.J., Mao, C.H., Dai, J.H., Kuang, Y.T., 2017. Sec-buzzer: cyber security emerging topic mining with open threat intelligence retrieval and timeline event annotation. Soft Computing 21 (11), 2883–2896.
Li, V.G., Dunn, M., Pearce, P., McCoy, D., Voelker, G.M., Savage, S., Levchenko, K., 2019. Reading the tea leaves: A comparative analysis of threat intelligence. In: Proceedings of the 28th USENIX Conference on Security Symposium. USENIX Association, USA, pp. 851–867. SEC'19
Liao, X., Yuan, K., Wang, X., Li, Z., Xing, L., Beyah, R., 2016. 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.
Long, Z., Tan, L., Zhou, S.. Collecting indicators of compromise from unstructured text of cybersecurity articles using neural-based sequence labelling.
Marsland, S., 2015. Machine learning: An algorithmic perspective. Chapman & Hall/CRC machine learning & pattern recognition series, Second edition CRC Press.
Mavroeidis, V., Bromander, S., 2017. 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.
MITRE. CAPEC - common attack pattern enumeration and classification (CAPEC). URL https://capec.mitre.org/index.html.
Mittal, S., Das, P.K., Mulwad, V., Joshi, A., Finin, T., 2016. 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.
Mittal, S., Joshi, A., Finin, T.. Cyber-all-intel: an AI for security related threat intelligence.
Mittal, S., Joshi, A., Finin, T.. Thinking, fast and slow: combining vector spaces and knowledge graphs.
Nebot, V., Rangel, F., Berlanga, R., Rosso, P., 2018. Identifying and classifying influencers in twitter only with textual information. In: Natural Language Processing and Information Systems, volume 10859. Springer, pp. 28–39.
OASIS Open. STIX: Cyber threat intelligence technical committee. URL https://oasis-open.github.io/cti-documentation/.
Pastor-Galindo, J., Nespoli, P., Gómez Mármol, F., Martínez Pérez, G., 2020. The not yet exploited goldmine of OSINT: opportunities, open challenges and future trends. IEEE Access 8, 10282–10304.
Pendlebury, F., Pierazzi, F., Jordaney, R., Kinder, J., Cavallaro, L., 2019. Tesseract: Eliminating experimental bias in malware classification across space and time. In: Proceedings of the 28th USENIX Conference on Security Symposium. USENIX Association, USA, pp. 729–746. SEC'19
Robertson, J., 2017. Darkweb cyber threat intelligence mining. Cambridge University Press.
Sabottke, C., Suciu, O., Dumitras, T., 2015. 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.
Sapienza, A., Bessi, A., Damodaran, S., Shakarian, P., Lerman, K., Ferrara, E., 2017. Early warnings of cyber threats in online discussions. In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, pp. 667–674. doi:10.1109/ICDMW.2017.94.
Sapienza, A., Ernala, S.K., Bessi, A., Lerman, K., Ferrara, E., 2018. 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.
Schaberreiter, T., Kupfersberger, V., Rantos, K., Spyros, A., Papanikolaou, A., Ilioudis, C., Quirchmayr, G., 2019. 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. ACM, pp. 1–10.
Scikit-learn: machine learning in python, URL https://scikit-learn.org/stable/.
Stone, M., 2021. The state of 0-day in-the-wild exploitation. USENIX Association. URL https://www.usenix.org/conference/enigma2021/presentation/stone
Subbian, K., Melville, P., 2011. Supervised rank aggregation for predicting influencers in twitter. In: 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing. IEEE, pp. 661–665. doi:10.1109/PASSAT/SocialCom.2011.167.
Tounsi, W., Rais, H., 2018. A survey on technical threat intelligence in the age of sophisticated cyber attacks. Computers & Security 72, 212–233.
Tundis, A., Mazurczyk, W., Mühlhäuser, M., 2018. A review of network vulnerabilities scanning tools: Types, capabilities and functioning, hamburg, germany, august 27–30, 2018. In: Proc.of the 13th International Conference on Availability, Reliability and Security (ARES 2018). ACM doi:10.1145/3230833.3233287.
Tundis, A., Mühlhäuser, M., 2017. A multi-language approach towards the identification of suspicious users on social networks. In: 2017 International Carnahan Conference on Security Technology (ICCST), Madrid, Spain, October 23–26, pp. 1–6. doi:10.1109/CCST.2017.8167794.
Twitter Inc.. Twitter API developer documentation. URL https://developer.twitter.com/.
Yang, J., Yu, M., Qin, H., Lu, M., Yang, C., 2019. A twitter data credibility framework-hurricane harvey as a use case. ISPRS Int J Geoinf 8 (3), 111. doi:10.3390/ijgi8030111. URL https://www.mdpi.com/2220-9964/8/3/111
Zhu, Z., Dumitras, T., 2018. 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.

Andrea Tundis is a Senior Researcher and his area of expertise are infrastructure protection, Internet organized crime and human safety. In 2014 he got a Ph.D. degree in Systems and Computer Science from the DIMES department at University of Calabria (Italy). He is currently working at Department of Computer Science at Technische Universität Darmstadt (TUDA) in Germany and member of the Telecooperation Lab (TK). He is involved in a Horizon 2020 European research project on organized cyber-crime and online crime detection by investigating on models and methods for the identification, prevention and response of Internet-based crimes.

Samuel Ruppert is a IT security expert at Deutsche Bahn Systel GmbH as well as scientific collaborator at Technische Universität Darmstadt (TUDA) within the Telecooperation Lab, in Germany. He received his M.Sc. degree in Informatics and IT Security in 2019, by focusing open source cyber threat intelligence sources detection, analysis and classification.

Max Mühlhäuser is a full professor at Technische Universität Darmstadt and head of Telecooperation Lab. He holds key positions in several large collaborative research centers and is leading the Doctoral School on Privacy and Trust for Mobile Users. He and his lab members conduct research on The Future Internet, Human Computer Interaction and Cybersecurity, Privacy & Trust. Max founded and managed industrial research centers, and worked as either professor or visiting professor at universities in Germany, the US, Canada, Australia, France, and Austria. He is a member of acatech, the German Academy of the Technical Sciences.