Corpus and Deep Learning Classifier for Collection of Cyber Threat Indicators in Twitter Stream

Vahid Behzadan, Carlos Aguirre, Avishek Bose and William Hsu
Department of Computer Science
Kansas State University
{*behzadan, caguirre, abose, bhsu*}*@ksu.edu*

Abstract

This paper presents a framework for detection and classification of cyber threat indicators in the Twitter stream. Contrary to the bulk of similar proposals that rely on manually-designed heuristics and keyword-based filtering of tweets, our framework provides a data-driven approach for modeling and classification of tweets that are related to cybersecurity events. We present a cascaded Convolutional Neural Network (CNN) architecture, comprised of a binary classifier for detection of cyber-related tweets, and a multi-class model for the classification of cyber-related tweets into multiple types of cyber threats. Furthermore, we present an open-source dataset of 21000 annotated cyber-related tweets to facilitate the validation and further research in this area.

1. Introduction

To keep pace with the growing complexity and frequency of cyber attacks, defensive operations are increasingly reliant on proactive measures. Such approaches require the timely, accurate, and actionable understanding of the threats that pose potential risks to protected systems. To meet this vital need, the paradigm of Cyber Threat Intelligence (CTI) has been introduced as a framework to facilitate the exploration, collection, and analysis of various sources of information on cyber threats.

An an important source of information, Open-Source Intelligence (OSINT) have proven to be a valuable resource for CTI. In particular, Twitter is deemed as a rich source of OSINT. The popularity of this medium among the cybersecurity community provides an environment for both the offensive and defensive practitioners to discuss, report, and advertise timely indicators of vulnerabilities, attacks, malware, and other types of cyber events that are of interest to CTI analysts. The value of Twitter with regards to CTI is well-demonstrated by the numerous initial reports of major cyber events, recent examples of which include disclosures of multiple 0-day Microsoft Windows vulnerabilities[1], user reports on DDoS attacks [1], and exposure of ransomware campaigns [2].

Over the recent years, the research on Twitter-based OSINT collection has led to the proposal of multiple frameworks (e.g., [3], [4], [5], [6], [7], [8], [9]) for detection and analysis of threat indicators in the Twitter stream. However, the majority of these proposals are heavily based on manual heuristics such as keyword lists for detecting and filtering tweets that are relevant to cybersecurity. This will inevitably lead to high false-positives in the detection of relevant tweets (e.g., filtering for the keyword "vulnerability" may result in storing a personal or spiritual tweet as one related to cybersecurity). Also, the flexible typography and the emergence of new terminology lead to the neglection of potentially valuable information in tweets. Furthermore, current state of the research in this area still lacks open-source dataset of manually annotated cyber-related tweets, which curtails further efforts to validate, compare, and extend current frameworks.

Utilization of OSINT in CTI, particularly via social informatics and text analytics, incur the challenges of document filtering and threat identification. In this work we describe the development of a social media test bed based on information extraction and machine learning for relevance filtering and classification of new intelligence with respect to defined threat categories. This test bed in turn is part of a data mining

1. https://www.zdnet.com/article/microsoft-windows-zero-day-disclosed-on-twitter-again/

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

pipeline and framework for threat intelligence, made extensible through the development of an annotation user interface and a flexible tag set which we apply to a corpus crawled from Twitter to demonstrate the effectiveness of the overall system. We present positive initial results using supervised inductive learning on the annotated corpus. We then review potential uses of both the test bed, and the annotation, machine learning, and OSINT data acquisition software components that produced this test bed, to open problems in threat intelligence, particularly those involving predictive analytics, empirical methods for natural language processing such as topic modeling and knowledge base population, and heterogeneous information network analysis.

Accordingly, the main contributions of this paper are:

1) Curation of an open-source dataset of 21000 manually annotated cyber-related tweets to facilitate further research on OSINT collection and analysis from Twitter,
2) Development of an open-source web application for annotation, exploration, and management of Twitter-based OSINT collections, and
3) Proposal and Validation of a data-driven approach to the detection and classification of cyber-related tweets based on Convolutional Neural Networks (CNNs).

The remainder of this paper is organized as follows: Section 2 provides an overview of the related literature. Section 3 documents the collection and annotation process of the cyber-related tweets dataset, followed by the details and evaluation of the proposed deep learning model for detection and classification of cyber-related tweets in the Twitter stream in Section 4. Finally, Section 5 concludes the paper with remarks on future directions of research.

2. Related Work

While the detection and classification of tweets has been widely explored in domains such as disaster response [10], crime prevention [11], and identification of cyber-bullies [12], the domain of CTI extraction from Twitter is lesser explored. Khandpur et al. [3] propose a framework to extract cyber threat and security information from the twitter data with the aim of identifying three types of threats and events, namely Distributed Denial of Service (DDoS) attacks, data breaches, and account hijacking. Their framework is comprised of three major components, labeled by the authors as target domain generation, dynamic typed query expansion, and event extraction. This approach is shown to be effective due to its exploitation of both syntactic and semantic analysis, as well as a dependency tree graph. However, this approach necessitates the continuous tracking of baselines and features for each type of threat. Furthermore, it demands a high computational overhead for generating and maintaining the targeted corpus domain of tweet text for query expansion. Additionally, this framework is unable to seamlessly extend to more categories of threats and events.

Another framework is proposed by Le Sceller et al. [4], which applies an unsupervised approach for detecting and categorizing cybersecurity events from tweets. Their proposed approach is based on a set of seed keywords specified for each level of the CTI taxonomy. Accordingly, [4] presents a method for expanding the set of seed keywords by identifying and appending new words with similar meanings in the context of word embeddings using a manually specified threshold in the cosine similarity distance between word vectors. This framework considers events as clusters of tweet texts generated via the TF-IDF method [REF]. However, this algorithm is prone to high false-positive rates due to the inadvertent biasing effects of the initial seed keywords. Also, a fixed, manually-specified threshold for annexation of new words proves to be inefficient in the effective selection of new keywords for a particular CTI-related event type.

In another approach proposed in [5], tweets are processed by the Security Vulnerability Concept Extractor (SVCE) [REF] which is trained on a dataset comprised of reports in the National Vulnerability Database to identify and tag the terms and concepts related to CTI, such as the means of attack, consequences of attack, and the affected software, hardware, and vendors. With such tags available, the concepts and entities extracted by SVCE are analyzed based on external publicly available semantic knowledge bases such as DBPedia, to further enrich their extracted data. This framework is developed for customer-based applications, and thus requires that the user specifies a target system profile comprised of information about installed software or hardware. Accordingly, an ontology is developed and used along with SWRL rules to address and prioritize time-sensitive CTI entries. The extracted and tagged CTIs are also converted to sets of RDF triple statements. The RDF linked data representation is stored in a knowledge base, thus allowing the alert system

to reason over the data. The limitations of SVCE in analyzing unofficial CTI-related text, as well as the reliance on hand-crafted rules, result in the inherent ineffectiveness of this framework in detecting novel threat types and indicators.

[6] proposes a framework that incorporates Named Entity Recognition (NER) and ontology-based techniques to classify tweets as CTI-related events or non-events. If a tweet is classified as relevant, this framework performs topic detection via cross-referencing of the NER results with external knowledge bases such as DBPedia. Furthermore, this work produces an annotated dataset of tweet CTI and event types using Wikipedia's Current Event Portal, as well as human input collected via Amazon Mechanical Turk. With this annotated dataset, the authors study the performance of various machine learning approaches such as Naive Bayes, Support Vector Machines (SVM), and Long-Short Term Memory (LSTM) recurrent neural network architectures, and report that the LSTM architecture with word embedding for feature representation produces the best results. They also demonstrate that the generic category of NER is helpful in the binary classification of relevance, whereas specific categories of NER are helpful in classifying the event type and categories. For topic identification, this work adopts the Pagerank algorithm to identify the closest topic in the relational graph of the tweet concept.

Lee et al. [7] focus on the detection of communities and influential user of Tweeter to prioritize CTI information via scoring the expertise of each user and community that produces CTI-related tweets. This framework is comprised of four components. The first component is a social media connector which connects and gathers data from the Twitter platform. The second component is a module for identifying and extending the list of experts to find emerging topics. Weight contribution and fitness calculation is referred as the third component for explaining the process to of each targeted expert, and how to mine valuable security-related information from the tweets posted by experts. Lastly, exploring then exploiting information from the expert pool is to recognize emerging threats with a LDA-based topic detection algorithm proposed. This method is dependent highly on expert identification and extracting information from them. Thus it can mislead to notify if the expert are not actually expert and threat indications are not sufficiently referred to by the experts.

[9] proposes a weakly supervised learning approach to train a model for extracting new categories of cybersecurity events by seeding a small number of positive event samples over a significant amount of unlabeled data. A learning objective has been employed here to regularize the label distribution over the unlabeled distribution towards user-provided expectation. This approach is heavily dependent on historical seed examples per event category. Also, this work fails to provide the details of matching named entities into an event category.

3. Data Collection and Annotation

To enable a supervised deep learning approach to the problem, a corpus of 21,000 tweets was curated directly from Twitter[2]. A custom stream listener made with Tweepy [13] was developed to listen to the live stream of tweets. A list of keywords was selected to pre-filter and narrow down the stream listener results. General words like "vulnerability" and "0day" were selected for their general relevance to CTI, while words related to specific types of threats like "DDoS", "SQL injection", "buffer overflow", among others, were selected to produce more targeted filtering. The tweet corpus was the result of a continuous stream listened over a four-day period. Because of the keyword selection, a considerable number of tweets were unrelated to CTI or to science and technology altogether, and hence it became apparent that a binary classification based on relevance to cyber threats was needed. Also, a classification over threat types would be of special interest as it could aid on threat event discovery.

3.1. Pre-Processing

During the tweet collection process, tweets were pre-labeled in terms of relevance and type to serve as suggestions in the annotation process with the goal of helping to speed up and increase the accuracy of manual annotation. For the first binary pre-labeling, the topic modeling API of IBM's Watson Natural Language Understating service [14] was used to recollect text classification into categories for the textual contents of each tweet. The five-level category hierarchy that Watson uses the assign categories to text was studied and relevant tags were extracted that had a connection to cybersecurity. The text category assignment was restricted to the top three categories

2. The dataset is available at https://github.com/behzadanksu/cybertweets

with highest confidence score. Thus, the binary pre-labeling was determined on the presence of relevant categories (e.g., computer networks, computer security, technology, etc) in the top three list that was assigned by Watson. Furthermore, the pre-labeling of types was performed by simple string matching on the pure tweet text. The types considered were: "vulnerability", "DDoS", "ransomware", "botnet", "data leak", "zero day" and "general".

3.2.Annotation

After the collection and pre-labeling of tweets, the annotation of all 21,000 tweets was performed by four human cybersecurity experts. The annotation of tweets with regards to both relevance and threat types was performed concurrently. To make the annotation of relevance easier for the annotators, a third subclass of "general or marketing" was added to disambiguate the nature of tweets that were related to cybersecurity but did not specifically concern any threats. To reduce the complexity of the annotation task, a user interface was created to speed up the annotation process and to aid annotators. In the user interface, a list of tweets was shown to the annotator, including the list of the categories that Watson classified the text of the tweet as, along with the corresponding confidence score. To classify for the type of threat, a drop-down list with the different threats was available to the annotator, which automatically updated the tweet type when changed. Also, for the binary classification, three color-coded buttons were available for "relevant", "not relevant" and "general or marketing", which automatically updated the tweet with a single click. With these features, the annotator could scroll down a list of tweets that showed their full text and Watsons NLP text categories, and annotate for both parameters with two clicks, thus increasing the annotation speed and efficacy. A sample of the annotation interface is illustrated in figure 1.

4. Classification Model

To increase the accuracy and extent of OSINT collection, we propose a cascade of two Convolutional Neural Network (CNN) models with identical architectures. The first CNN is trained to classify each tweet as relevant or irrelevant to cybersecurity. If the tweet is classified as relevant, then it is passed to the second CNN to be classified as one of 8 different types, namely: vulnerability, DDoS, data leak, ransomware,

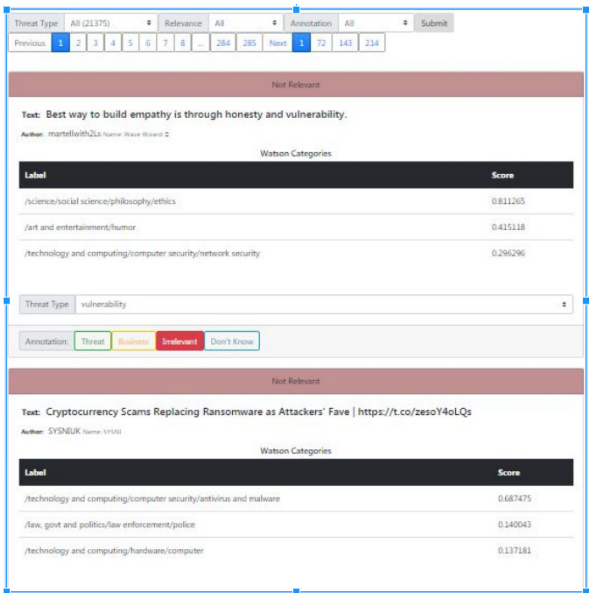

Figure 1. The annotation interface

0-day, and marketing/general. The training and validation sets are provided from the annotated dataset with a ratio of $9:1$, respectively. To encode the samples for these models, the pre-processed text of each tweet is mapped to a 150-dimensional vector of word embeddings [15]. The remainder of this section provide further details on the pre-processing steps and the model's architecture.

4.1.Pre-Processing

Before transforming into word embedding vectors, the full text of each tweet is pre-processed to prepare a more coherent representation of the entire dataset with few redundancies. As illustrated in figure 2, these steps are: (1) conversion of all characters of the tweet to lower case, (2) tokenize the text according to white-space separations, (3) remove tokens that are not encoded in ASCII, (4) remove punctuations from each token, (5) remove tokens that are not comprised of alpha-numeric characters, (6) substitute digits with word representations (e.g., $4 \rightarrow$ four), (7) remove stop words, (8) stem tokens. It is noteworthy that steps (7) and (8) utilize the functionalities available in the Natural Language ToolKit (NLTK) libraries [16].

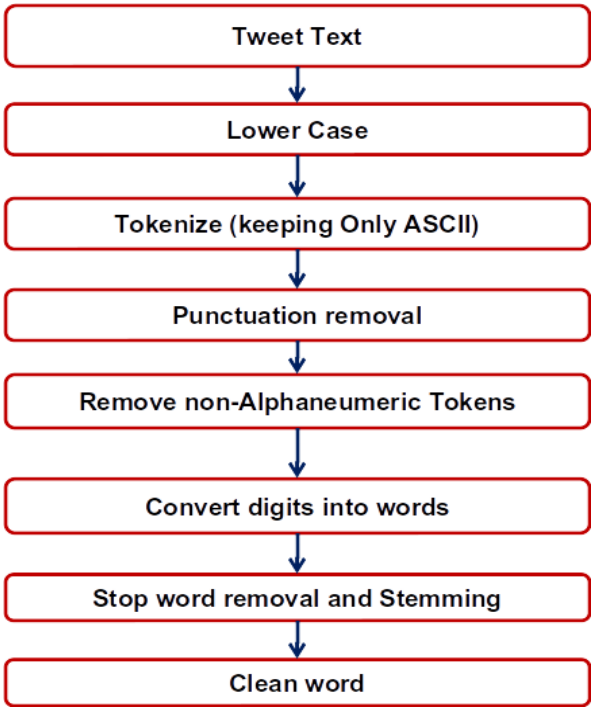

Figure 2. Pre-Processing Steps

4.2.Model Architecture and Configuration

As mentioned, the input to the proposed CNN model is a 150-dimensional word embedding vector. This is followed by a convolutional layer with 32 filters as parallel fields for processing words, and a kernel size of 8 with a rectified linear (reLu) activation function. Next is a pooling layer of size 2 to reduce the output of the CNN layer by half, which is then flattened to a 2-dimensional vector, representing the features extracted by CNN. The proceeding layer is a standard Multi-Layer Perceptron (MLP) comprised of 30 reLu cells to interpret the CNN features. Finally, a sigmoid activation function is used in the output layer to return the class with the highest probability.

To account for the imbalance caused by the larger number of "relevant" samples to "irrelevant" samples, the first classifier uses weighted classes of ratio $1000 : 75$, respectively. The training process uses the Adam optimizer [17] to minimize the categorical cross-entropy loss function [18].

4.3.Results

As depicted in Table 1, the binary classifier performs with a mean accuracy of 94.72%. On the other hand, the multi-class classifier demonstrates a slightly worse performance. This can be attributed to the lower number of samples available for each type, in comparison to that of the relevance vs. irrelevance. Unfortunately, due to the unavailability of datasets, these results cannot be compared with those of the previous works.

Table 1. Performance of Detection and Classification Models

	CTI Relevance	Threat Type
Mean Accuracy	94.72%	87.56%
Mean Recall	94.57%	85.48%
Mean F1	94.62%	81.99%

5. Conclusion

Preliminary results on the test bed task of cyber-threat relevance (a binary classification or concept learning task) are promising, as the mean F1 score of nearly 0.95 indicates good potential for pre-trained CNNs as a Bloom filter in a big data analytics pipeline. Results on the multiclass classification task of threat type identification leave much more room for improvement, with a mean F1 score of just under 0.82. Training on larger corpora, multi-annotator agreement, and weight regularization for feature extraction and selection are methods that may improve this score, particularly by raising the cross-validated mean average precision.

In current and future work, we maintain our focus on *event detection and tracking* [19] of cyber-threats, using social network analysis [20] for prioritization of threat indicators, and multi-source intelligence fusion methods [21] and analytics for monitoring of active threats.

References

[1] C. Sabottke, O. Suciu, and T. Dumitras, "Vulnerability disclosure in the age of social media: Exploiting twitter for predicting real-world exploits." in *USENIX Security Symposium*, 2015, pp. 1041–1056.

[2] A. Sapienza, A. Bessi, S. Damodaran, P. Shakarian, K. Lerman, and E. Ferrara, "Early warnings of cyber threats in online discussions," in *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*. IEEE, 2017, pp. 667–674.

[3] R. P. Khandpur, T. Ji, S. Jan, G. Wang, C.-T. Lu, and N. Ramakrishnan, "Crowdsourcing cybersecurity: Cyber attack detection using social media," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 1049–1057.

[4] Q. Le Sceller, E. B. Karbab, M. Debbabi, and F. Iqbal, "Sonar: Automatic detection of cyber security events over the twitter stream," in *Proceedings of the 12th International Conference on Availability, Reliability and Security*. ACM, 2017, p. 23.

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

[6] A. Edouard, "Event detection and analysis on short text messages," Ph.D. dissertation, Université Côte d'Azur, 2017.

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

[8] A. Sapienza, S. K. Ernala, A. Bessi, K. Lerman, and E. Ferrara, "Discover: Mining online chatter for emerging cyber threats," in *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 2018, pp. 983–990.

[9] A. Ritter, E. Wright, W. Casey, and T. Mitchell, "Weakly supervised extraction of computer security events from twitter," in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 896–905.

[10] Z. Ashktorab, C. Brown, M. Nandi, and A. Culotta, "Tweedr: Mining twitter to inform disaster response." in *ISCRAM*, 2014.

[11] T. Mackey, J. Kalyanam, J. Klugman, E. Kuzmenko, and R. Gupta, "Solution to detect, classify, and report illicit online marketing and sales of controlled substances via twitter: Using machine learning and web forensics to combat digital opioid access," *Journal of medical Internet research*, vol. 20, no. 4, 2018.

[12] P. Galán-García, J. G. d. l. Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," *Logic Journal of the IGPL*, vol. 24, no. 1, pp. 42–53, 2016.

[13] J. Roesslein, "tweepy documentation," *Online] http://tweepy. readthedocs. io/en/v3*, vol. 5, 2009.

[14] Y. Chen, J. E. Argentinis, and G. Weber, "Ibm watson: how cognitive computing can be applied to big data challenges in life sciences research," *Clinical therapeutics*, vol. 38, no. 4, pp. 688–701, 2016.

[15] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[16] S. Bird and E. Loper, "Nltk: the natural language toolkit," in *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, 2004, p. 31.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[19] W. Elshamy and W. H. Hsu, "Continuous-time infinite dynamic topic models: The dim sum process for simultaneous topic enumeration and formation," in *Emerging Methods in Predictive Analytics: Risk Management and Decision-Making*, W. H. Hsu, Ed. Hershey, PA, USA: IGI Global, 2014, pp. 187–222.

[20] M. Yang, W. H. Hsu, and S. Kallumadi, "Predictive analytics of social networks: A survey of tasks and techniques," in *Emerging Methods in Predictive Analytics: Risk Management and Decision-Making*, W. H. Hsu, Ed. Hershey, PA, USA: IGI Global, 2014, pp. 297–333.

[21] A. Danyluk, T. Fawcett, and F. Provost, Eds., *Heterogeneous Time Series Learning for Crisis Monitoring*. AAAI Press, 1998.