POSTER: Analysis and Parsing of Unstructured Cyber-Security Incident Data

Armando J. Ochoa
aocho032@fiu.edu
Florida International University
Miami, FL

Mark A. Finlayson
markaf@fiu.edu
Florida International University
Miami, FL

ABSTRACT

The latest threat intelligence platforms use structured protocols to share and analyze cyber-security data. However, most of this data is reported to the platform in the form of unstructured text such as social media posts, emails, and news articles, which then require manual conversion to structured form. In order to bridge the gap between unstructured and structured data, we propose to implement a natural-language-processing-(NLP)-based information extraction (IE) system that takes texts within the cyber-security domain and parses them into structured format. Our approach targets the VERIS format and makes use of the VERIS Community Database as a source of unstructured texts—primarily consisting of news articles–and their structured counterparts (VERIS reports). We propose first to use a supervised machine learning (ML) classifier to discriminate between cyber-related and non-cyber-related texts, and then to use ML classifiers decide which VERIS parameters are relevant in a given text. Then, we propose to use NLP and IE techniques to extract tuples of grammatically co-dependent words. Finally, these tuples will be passed to a domain- and field-specific IE components to fill in different fields of an output VERIS report.

KEYWORDS

Cyber-security, Information Extraction, Natural Language Processing, VERIS

ACM Reference Format:
Armando J. Ochoa and Mark A. Finlayson. 2019. POSTER: Analysis and Parsing of Unstructured Cyber-Security Incident Data. In *WiSec '19: ACM Conference on Security and Privacy in Wireless and Mobile Networks, May 15–17, 2019, Miami, FL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3317549.3326324

Cyber-security professionals at every level rely on up-to-date and accurate information about potential threats and actors, as well as the tools they might use and exploits they might targets. The number of these potential threats, actors, tools, and exploits, as well as the information required to understand them and track them, has grown exponentially in the past few decades. Threat intelligence platforms (TIPs) have been established to share and analyze cyber-security incident data, and they use structured protocols (such as STIX2 or VERIS) to provide consumable feeds to downstream analysts. However, most cyber-security incidents are reported and discussed using unstructured texts such as blog posts, news articles, advisories, social media activity, email chains, etc. Due to the unstructured nature of these natural language texts, machines cannot easily consume and process them, reducing the amount of information to which threat intelligence platforms and analysts have access. To address this problem, we propose implementing a system that uses natural language processing (NLP) and information extraction (IE) to transform unstructured texts in the cyber-security domain into a structured format.

Our system is currently mostly at the design stage, with only small parts implemented. Therefore, in this extended abstract we describe the proposed design and the various components that we believe will be required to effect a functional text-to-structured cyber-incident parsing system.

The system we propose will target the VERIS report format [7]. The motivation for this choice is two-fold. First, VERIS is an established structured format that specifically models cyber-related events. Second, the VERIS Community Database (VCDB) provides a ready-made dataset of VERIS reports paired with their source news articles on which to develop, train, and test our proposed system. A VERIS report in the VCDB describes various aspects of a single cyber-security incident (breaches, infiltrations, etc.) in a large number of special purpose fields, and also contains references to news articles that describe the incident in unstructured text.

Because the ultimate goal is to build a system that monitors a news feed for relevant articles, we propose that the system begin first with a classifier that can discriminate between cyber- and non-cyber-related documents. In preliminary experiments we trained a support vector machine (SVM) binary model using libsvm [3] on simple bag of words and *tf-idf* features. We will use a second set of news articles from the Open American National Corpus (OANC) [4] *journal* domain as negative examples for testing and training.

Once cyber-relevant documents are identified, we propose to run each document through a set of binary classifiers that have been trained for each parameter of the VERIS format to decide whether or not that parameter is relevant to the text. Positive and negative examples for these binary models will be obtained from the existing VERIS reports.

The actual values for each parameter will be obtained using deeper NLP and IE techniques. We propose, first, to preprocess each text with the Stanford CoreNLP [6] toolkit, including the steps such as tokenization, sentence splitting, lemmatization, part-of-speech tagging, and dependency parsing.

Of particular importance to VERIS reports will be Named Entity Recognition (NER) and Relation Extraction (RE). Cyber-domain-specific NER and RE systems have been explored in prior work [2, 5].

WiSec '19, May 15–17, 2019, Miami, FL, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6726-4/19/05.
https://doi.org/10.1145/3317549.3326324

Inspired by these approaches, we will build NER and RE components specific to each VERIS field. These NER and RE components will use both manually curated cyber-incident keywords as well as keywords identified using *tf-idf* and other mutual information measures [1]. Our assumption is one of the most important sets of features will be common terms (*leak*, *infiltration*, *disclosure*, etc.) that are strongly associated with relevant pieces of information such as *what* was disclosed or *who* infiltrated. Feature weights will be calibrated by comparison with non-cyber-related corpora (e.g., the OANC) in order to eliminate common and irrelevant terms. In support of RE specifically, we will also generate *N*-tuples using relevant keywords and all their grammatically dependent words using dependency parsing. The relevant words would include entities identified by the NER systems, and class-relevant keywords described above, but excluding irrelevant relations such as those involving punctuation. Whenever applicable, the NER and RE classifiers will be trained on data extracted from the VCDB itself.

REFERENCES

[1] Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management* 39, 1 (2003), 45–65.

[2] Robert A Bridges, Corinne L Jones, Michael D Iannacone, Kelly M Testa, and John R Goodall. 2013. Automatic labeling for entity extraction in cyber security. *arXiv preprint arXiv:1308.4941* (2013).

[3] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27. Issue 3. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[4] Nancy Ide and Catherine Macleod. 2001. The American National Corpus: A standardized resource of American English. In *Proceedings of 1st Corpus Linguistics Conference (CL2001)*. Lancaster, UK, 1–7.

[5] Corinne L Jones, Robert A Bridges, Kelly MT Huffer, and John R Goodall. 2015. Towards a relation extraction framework for cyber-security concepts. In *Proceedings of the 10th Annual Cyber and Information Security Research Conference*. Oak Ridge, TN, Article No. 11.

[6] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), System Demonstrations*. Baltimore, MD, 55–60. https://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf

[7] Verizon Security Research & Cyber Intelligence Center. [n. d.]. The VERIS Framework. http://veriscommunity.net/ Retrieved on April 8, 2019.