Analysis and Parsing of Unstructured Cyber-Security Data

Armando J. Ochoa* & Mark A. Finlayson

Internally, threat intelligence platforms use structured protocols (such as STIX2 or VERIS) to share and analyze cyber-security data but cyber-security-related events are usually reported and talked about using free-form texts such as blog posts, social media activity, and news articles. Due to the unstructured nature of these natural language texts, machines cannot easily consume and process them, which reduces how much information analysts and threat intelligence platforms have access to. To solve this problem, we propose implementing an Information Extraction System that takes unstructured texts within the cyber-security domain and processes and parses them into a structured format. We will create a pipeline which consumes free-from text articles taken from the VERIS Community Database and process them to create VERIS-sytle JSON reports. The Stanford CoreNLP toolkit will provide parsing, tokenization, and part-of-speech analysis to prepare the text for more complex information extraction techniques. Named Entity Recognition and Relationship Extraction of different levels, from regex to statistical models, together with rule-based analysis, are used to parse out the actors and events which contain cyber-related information and are relevant to the VERIS model. In order to complement the pipeline, we will also create a working set of annotated free-form texts out of a subset of the VERIS Community Database. To our knowledge, this Information Extraction system is the first one to be directly designed for the Cyber-Security Domain and the first one to leverage the VERIS format and the VERIS Community Database. This system would help bridge the gap between structured and unstructured data within the cybersecurity domain, allowing security specialist to easily consume the plethora of cyber-security-related data that is freely accessible over the internet.