Text Mining for Factors

TREO Talk Paper

Andrés Díaz López
Arizona State University
andres.diaz.lopez@asu.edu

Yuan Xue
Penn State Berks
yxx@psu.edu

Xiang Guo
Missouri State University
xguooffice@gmail.com

Abstract

This research aims to automate the time-consuming activity of literature reviews for the collection of indicators used in the assessment of constructs. The study contends that extraction of those potential indicators can be done through text mining techniques developed under specific semantic rules. Authors show their text mining semantic rules for the extraction of potential Digital Forensic Readiness (DFR) indicators from the literature. DFR is a construct of increasing interest in cybersecurity for which no commonly accepted assessment framework exists (Díaz López 2017), making it a perfect study case.

Authors extracted a raw list of 1,096 unique phrases that indicate DFR from 77 academic articles. Many of these phrases are not a measurable factor. That is, they do not propose the presence of a measurable condition that, at a lower or higher extent, affects the DFR status. Rather, they indicate an aspect or a context to take into account in the assessment of DFR. These can be considered dimensions of DFR. Authors separated these 1,096 phrases between those which could be measurable factors and those which indicated dimensions. This process required the revision of each phrase by two of the researchers, first, independently, and later, together, in order to settle disagreements in the classification. This classification resulted in 181 phrases classified as indicators of dimensions and 915 classified as indicators of factors. This process was not only time-consuming, but it might also be considered subjective in the extraction of potential indicators of the construct, as well as in their classification as measurable factors.

However, once decisions have to be made over ambiguous phrases. It became clear that the human process must follow semantic and ontological rules in order to find agreement on the validity of the phrase as an indicator. Moreover, if this observation is correct, it should be possible to build domain-specific semantic-ontological rules for the automatic extraction of potential indicators of a construct from the related academic literature. This means that we can build a semantically-ontologically-based text mining algorithm to help us extract and classify DFR indicators from the 77 papers reviewed, and then compare its performance with the human extraction and classification already done.

Authors developed an approach using the semantic function of words as entities or conditions and implemented it in a text mining algorithm in order to distinguish those phrases indicating factors from those indicating dimensions. The algorithm classification was contrasted with the original human classification and resulted in measures of 0.5 accuracy, 0.5 error, 0.9 precision, and 0.43 recall. This means that our algorithm is good at distinguishing factors from dimensions, but not good enough at detecting all potential factors. Further analysis of the semantic-ontological nature of our phrases is needed in order to refine the algorithm. In this process, the feedback from fellow researchers would be highly appreciated. A good performance of this algorithm can substantially reduce researchers work in the identification of DFR factors. Furthermore, semantic-ontological rules discovered in this process can be generalized and implemented in the examination of other constructs.

References

Díaz López. 2017. “Event-Based Assessment of Cyber Security and Digital Forensic Readiness,” Proceedings of the Americas Conference in Information Systems, Boston, August, 2017