Skysong Innovations
Advancing the Arizona State University Knowledge Enterprise

1475 N. Scottsdale Road, Suite 200
Scottsdale, AZ 85287-3538
Phone: 480 884 1996 Fax: 480 884 1984

Case ID:M18-191P
Published: 3/17/2022

Inventors

Revanth Patil
Paulo Shakarian
Ashkan Aleali
Ericsson Marin

Contact

Shen Yan
shen.yan@skysonginnovations.com

Automatic Assignment of Labels to Discussion Topics Seen in Dark Web Forums

-Background The ability to traverse the internet with complete anonymity provides online platforms for illegal activities such as credit card fraud, identity theft, and leaks of sensitive information. One of the most prevalent cyber environments that has emerged over the last decade and enabled criminality are dark web forums, since they include encryption technology to prevent monitoring and also provide protection from unauthorized users. Considering the enormity of the data in those environments, there is an impending need to provide for security researchers a granular, structural, and interdependent classification of the available information. Current technologies use learning models and techniques that do not address the issues of labeled data scarcity, nor do they address imbalanced data classes in the training set. Training sets are labeled by hand, which is a time-consuming and typically unscalable process. Thus, a more thorough and versatile learning method is desired. Invention Description Researchers at Arizona State University have developed an intelligent system capable of classifying the information extracted from dark web forums in a hierarchical structure of tags. To solve class imbalance and scarcity of labeled data problems, a semi-supervised model was created based on elastic search (ES) document relevance score. Ensuring hierarchical integrity constraints improved the F1 score by 11.9% over standard supervised learning, while the ES-based semi-supervised learning model outperformed other models in terms of precision (78.4%) score while maintaining comparable recall (21%) score.

The system crawls various dark web sites and extracts important information from HTML pages, storing it in a database. Specialized crawlers automatically connect and authenticate the dark web sites through Tor. Important information, such as discussion topics and user-related information, is then parsed from these sites and stored on a database as well as on an elastic search data store. Data preprocessing and feature extraction is then performed over the forum discussion topics.

This technology is covered by U.S. Pat. No. 11,275,900.

Potential Applications • Cybersecurity • Dark web monitoring • Fraud detection Faculty Profile of Professor Paulo Shakarian