Context-Aware Feature Vectors in Dark Web Page Classification

Sander Brinkhuijsen[1,2],
Supervisors: Romana Pernisch[1,3] Eljo Haspels[2], and Mark van Staalduinen[2]

[1] Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
[2] CFLW Cyber Strategies, The Hague, the Netherlands
[3] Discovery Lab, Elsevier, Amsterdam, the Netherlands
sander.brinkhuijsen@cflw.com, r.pernisch@vu.nl, eljo.haspels@cflw.com,
mark.vanstaalduinen@cflw.com

Keywords: Dark Web · Web Page Classification · Vectorization.

Introduction. Privacy and anonymity are the most important virtues in the Dark Web which are ensured using encrypted communication over private or peer-to-peer networks [10]. However, Dark Web services are two sided coins. On the one hand, the Dark Web facilitates journalists and whistle-blowers with a tool to circumvent censorship and share sensitive information [3, 7]. On the other hand, it provides excellent cover for criminal activities as it is difficult for law enforcement agencies to establish their identity [2].

Monitoring the Dark Web for illegal activities is a time-consuming activity for law enforcement agents. CFLW Cyber Strategies (CFLW)[4], a company focused on providing intelligence services to increase cyberspace safety, introduced the Dark Web Monitor (DWM) to tackle this issue. The DWM is an open-source intelligence repository that provides insights into criminal and fraudulent activities facilitated on the Dark Web. It contains information about more than 1.5 million Dark Web domains. Each domain is assigned a set of (multiple) labels to describe its content, which are then used by investigators to filter out domains containing irrelevant topics. Currently, all domains in the DWM are annotated by a domain expert, which is time-consuming, error-prone and troublesome because their expertise can be used more effectively elsewhere.

This thesis proposes a classification pipeline for automated labelling of web pages from the Dark Web similar to the basic pipeline introduced by Al Nabki et al. [1]. Our research focuses on converting text into a numerical representation which is required for classification. A literature survey on web classification on the clear web found that traditional vectorization techniques, such as TF-IDF, are often used and more modern techniques are under-explored [5]. Literature targeted towards classifying Dark Web pages [9, 8, 4, 6, 1] shows the same characteristic. To modernize the field of web classification, this thesis investigates the *effect of context-awareness in feature vectors on web page classification by comparing TF-IDF, Word2Vec and SBERT as text vectorization methods.*

[4] https://cflw.com/

Where TF-IDF has no context-awareness, Word2Vec has intermediate context-awareness, and SBERT has advanced context-awareness.

Approach. The dataset used during the experiments is a subset of the DWM, it contains 9408 English domains covering twelve different labels. Six of them relate to the abuse type (Drugs/Narcotics, Sexual Abuse, Goods/Services, Financial Crime, Cybercrime, Violent Crime) and the other six relate to the provided service (Shop, File Sharing, Service Provider, Market, Index, Messaging Service). These labels are not mutually exclusive, meaning that each domain can be assigned multiple labels. We split the data into a train (70%), validation (20%) and test set (10%). Three different models, one for each vectorization method, have been created by means of the following four steps in the classification pipeline. First, the **Content Extraction** step extracts text from the raw HTML source code. The **Preprocessing** step then focuses on cleaning the text to reduce noise and decrease variability. The extend to which a text is preprocessed depends on the used vectorization method. Therefore, the amount of preprocessing decreases as context-awareness increases. Third, the **Vectorization** step converts text into a numerical representation. Instead of training Word2Vec and SBERT from scratch, which is a time consuming task, we used pretrained models. Gensim's *word2vec-google-news-300*[5] model was used for Word2Vec, and *paraphrase-MiniLM-L6-v2*[6] for SBERT. Lastly, the **Classification** step uses a single layer neural network to predict the set of labels which is iteratively trained on the training and validation set.

Findings. TF-IDF, Word2Vec, and SBERT achieve a macro average F1-score of 0.94, 0.88 and 0.95 respectively. Inspecting the performance per label shows that some labels are more difficult to predict than others as their performance is relatively low (0.75 average F1-score) in comparison to the overall performance. This decreased performance is most likely caused by a lack of sufficient samples, removal of information during the preprocessing step, and/or vague label definition. However, even for these labels, there is minimal difference in performance between TF-IDF and SBERT. On the one hand comparing operational performance, we found that the limited number of required preprocessing steps allows SBERT to have the lowest average processing time per sample. On the other hand, inspecting the storage requirements showed that TF-IDF has the lowest storage demands. Our findings show no indication that increased context-awareness leads to higher quality feature vectors for the task of web classification, since the context-unaware method (TF-IDF) and the context-aware method (SBERT) attain similar predictive performances. Nonetheless, we recommend CFLW to use SBERT over TF-IDF due to its superior operational performance (time). To conclude, this thesis modernizes the Dark Web classification task by comparing a traditional, context-unaware vectorization method (TF-IDF) to modern NLP techniques, Word2Vec and SBERT.

[5] https://radimrehurek.com/gensim/models/word2vec.html

[6] https://www.sbert.net/docs/pretrained_models.html

Acknowledgements The work described in this paper is performed in the H2020 project STARLIGHT (“Sustainable Autonomy and Resilience for LEAs using AI against High priority Threats”). This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 101021797.

Bibliography

[1] Al Nabki, M.W., Fidalgo, E., Alegre, E., de Paz, I.: Classifying Illegal Activities on Tor Network Based on Web Textual Contents. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 35–43, ACL, Valencia, Spain (2017), https://doi.org/10.18653/v1/E17-1004

[2] Chertoff, M.: A public policy perspective of the Dark Web. Journal of Cyber Policy **2**(1), 26–38 (Jan 2017), ISSN 2373-8871, 2373-8898, https://doi.org/10.1080/23738871.2017.1298643

[3] Finklea, K.: Specialist in Domestic Security. Dark Web p. 19 (Mar 2017)

[4] Ghosh, S., Das, A., Porras, P., Yegneswaran, V., Gehani, A.: Automated Categorization of Onion Sites for Analyzing the Darkweb Ecosystem. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1793–1802, ACM, Halifax NS Canada (Aug 2017), ISBN 978-1-4503-4887-4, https://doi.org/10.1145/3097983.3098193

[5] Hashemi, M.: Web page classification: a survey of perspectives, gaps, and future directions. Multimedia Tools and Applications **79**(17-18), 11921–11945 (May 2020), ISSN 1380-7501, 1573-7721, https://doi.org/10.1007/s11042-019-08373-8

[6] He, S., He, Y., Li, M.: Classification of Illegal Activities on the Dark Web. In: Proceedings of the 2019 2nd International Conference on Information Science and Systems, pp. 73–78, ACM, Tokyo Japan (Mar 2019), ISBN 978-1-4503-6103-3, https://doi.org/10.1145/3322645.3322691

[7] Rudesill, D.S., Caverlee, J., Sui, D.: The Deep Web and the Darknet: A Look Inside the Internet's Massive Black Box. SSRN Electronic Journal (2015), ISSN 1556-5068, https://doi.org/10.2139/ssrn.2676615

[8] Sabbah, T., Selamat, A., Selamat, M.H., Ibrahim, R., Fujita, H.: Hybridized term-weighting method for Dark Web classification. Neurocomputing **173**, 1908–1926 (Jan 2016), ISSN 09252312, https://doi.org/10.1016/j.neucom.2015.09.063

[9] Saini, J.K., Bansal, D.: A Comparative Study and Automated Detection of Illegal Weapon Procurement over Dark Web. Cybernetics and Systems **50**(5), 405–416 (Jul 2019), ISSN 0196-9722, 1087-6553, https://doi.org/10.1080/01969722.2018.1553591

[10] Zhang, H., Zou, F.: A Survey of the Dark Web and Dark Market Research. In: 2020 IEEE 6th International Conference on Computer and Communications (ICCC), pp. 1694–1705 (Dec 2020), https://doi.org/10.1109/ICCC51575.2020.9345271