DarkNER: A Platform for Named Entity Recognition in Tor Darknet

Mhd Wesam Al-Nabki
Dept. IESA.
Universidad de León
Researcher at INCIBE
mnab@unileon.es

Eduardo Fidalgo
Dept. IESA.
Universidad de León
Researcher at INCIBE
eduardo.fidalgo@unileon.es

Javier Velasco Mata
Dept. IESA.
Universidad de León
Researcher at INCIBE
jvelm@unileon.es

***Abstract*—In this paper, we introduce DarkNER, an application of Named Entity Recognition (NER) based on neural networks to identify six categories of named entities: Location, Person, Products, Corporation, Group, and Creative-work, in onion domains on the Tor network. The presented NER model is trained on the W-NUT-2017 dataset and tested on manually tagged samples of Tor hidden services. The experiments show the adaptability and effectiveness of neural networks models in detecting new textual entities, such as drugs names and weapons brands. The proposed application could help the authorities in filtering and monitoring the contents of the Tor domains.**

***Index Terms*—Named Entity Recognition, Darknet, Tor Network, Cybersecurity, Hidden Services**

Contribution Type: *Ongoing research*

Named Entity Recognition (NER) aims to identify different types of entities, such as a location, people names or products, within a given text. Those entities can be useful for several Natural Language Processing (NLP) tasks such as web contents filtering and monitoring, entity-based trend detection, and content mining [1], [2]. Recently, The Onion Router (Tor) network has become a safe shelter for practicing suspicious activities on the Internet, like drugs trading, weapons markets, or child pornography forums, far from the authorities' monitoring tools [2]. The conventional methods for monitoring the online textual contents, such as using predefined lists of keywords, allow to detect those keywords only, it is hard to maintain or keep it updated with emerging terms. Moreover, the problem becomes more complicated when those lists need to be created and updated in several languages. Hereafter, it would be useful to build an automatic system to recognize textual entities, with the capability to adapt to the dynamic content of the Tor domains.

Building a NER system is challenging due to the limited amount of supervised training samples and the possibility of having multiple meanings for a given word. Besides, the quality of the input text has a significant impact on the performance of the system . For example, a well-structured text quoted from a newspaper, where the capital letters and the punctuation marks are carefully reviewed, would be easier to understand rather than a short text written in slang words that may contain syntax mistakes.

In this paper, we present DarkNER, a platform to detect six categories of named entities (NE): Locations, Person, Creative-work, Group, Product, and Corporation, in the hidden services of the Tor network. In particular, we used the neural network model proposed by Aguilar et al. [8], since achieves the state of the art performance on W-NUT-2017 dataset[1]. To the best of our knowledge, the W-NUT-2017 is the most recent dataset for NE for noisy user-generated text.

Thanks to the experience we earned during labeling Darknet Usage Text Addresses (DUTA) dataset [3], we observed that the W-NUT-2017 dataset mimics the nature of the contents of the Tor network domains in terms of the quality of the text. Both datasets hold noisy user-generated text, which is rich with slang words, acronyms, abbreviations, together with the presence of emerging terms that have never been seen before. We propose to use the trained NER model of Aguilar et al. to detect NE on onion domains sampled from DUTA dataset.

The rest of the paper is organized as follows: Section I presents the related work. Then, Section II introduces the used neural network structure. After that, we explore the conducted experiments on DUTA samples in Section III. Finally, Section IV presents the conclusions by pointing out to our ongoing research on the field.

I. Related Work

The NER task has been a hot research topic for a long time. Before the rise of the deep learning techniques, the proposed methods mainly depended on manually extracted features from the input text. McCallum et al. [4] used handcrafted features, such as words prefix or suffix, and capital letters. However, the automatic feature extraction carried out using deep learning has pushed the state-of-the-art score strongly in NER systems. Lample et al. [5] proposed a neural network model with F1 score of 90.94% on Conll2003[2] dataset. Ma et al. [6] designed a model similar to Lample, but they used a Convolutional Neural Network (CNN) instead of Bi-LSTM for the characters sequences and had an F1 score of 91.21%. Although the neural network models have a high F1 score, the performance drops sharply in the case of the user-generated text like users' tweets on Twitter. Von Däniken et al. [7] used Transfer Learning (TL) and achieved F1 score of 40.78%. Aguilar et al. [8] presented a neural network model and trained it over the W-NUT-2017 dataset with F1 score of 41.86%.

II. Methodology

In this section, we describe briefly the neural network model proposed by Aguilar et al. [8]. The model extracts

[1]http://noisy-text.github.io/2017/emerging-rare-entities.html
[2]https://www.clips.uantwerpen.be/conll2003/ner

features from (i) word characters: an orthographic encoder is used to represent the characters. The encoded characters are embedded into a $\mathbb{R}^{d\times t}$ embedding space, where d is the dimension of the features per character and t denotes the word length threshold. Then, the result is passed into 2-stacked convolutional layers with a global average pooling. (ii) Word context: each word is represented using two codifications. First, pre-trained words embedding to capture latent semantics of words. Second, an embedding for the part-of-speech (POS) tags that were generated using the CMU Twitter POS tagger[3]. These embeddings are concatenated to form the final representation of the input word and passed into a Bidirectional Long Short-Term Memory (Bi-LSTM). (iii) A gazetteer, an external resource of knowledge. The gazetteer vector of a single word is a binary vector of n dimensions whereas n refers to the number of the categories in the dataset, i.e. 6 dimensions in the W-NUT-2017 dataset. The length of the gazetteer is equal to the size of the dataset's vocabulary. Next, the extracted features are fed into a multi-task network. The first task has a sigmoid activation function to identify whether the input token is an entity or not, while the second one has a softmax activation function to decide the category of the tag. Finally, the model is attached with a Conditional Random Field (CRF) to account for the sequential constraints in the input text.

III. Experimental Results

A. Tor Domains Entities Recognition

The W-NUT-2017 dataset has six types of NE that are encoded using: Begin, Inside, and Outside (BIO) tags such that the *B* in BIO refers to the beginning of a tag, the *I* refers to inside of a tag, and the *O* refers to non-entity words. Since there is no training dataset for the Tor hidden services, we used the W-NUT-2017 dataset for training. Later, to test the performance of the trained model, we manually labeled the NE tags of 15 onion domains which were randomly sampled from the categories Drugs and Violence in DUTA dataset. We found that the trained model can detect drugs names and weapons brands as *Products*, marketplaces names as a *Corporation*, names of cities and countries as *Location*, and people names as *Person*. Table I reports the performance of the model in terms of Precision, Recall, and their harmonic mean F1 score measures along with examples of the recognized entities per category.

TABLE I
Examples of the recognized NE in 15 samples onion domains along with precision/ recall and their harmonic mean F1 score

Categories	% Precision	% Recall	% F1 Score	NE Examples
Corporation	35.19	17.76	23.60	Alpha Pharma, Heckler & Koch
Creative-work	37.36	12.01	18.18	Danaucolt Ghost
Group	49.43	21.08	29.55	American brands
Location	53.80	52.94	53.37	Barcelona, Amsterdam
Person	68.35	51.37	58.65	Alex Grey, Vin Mariani
Product	56.48	20.85	30.46	marijuana, Purple Kush, S&W 10mm, Ruger M77
Average	56.72	30.47	39.65	

The results show that the system is capable of detecting NE with a low recall but with high precision, relatively. The decrease in the recall value could reflect the difficulty of recognizing emerging or rare terms in the test set. The model of Aguilar et al. depended on an external resource of knowledge that was built manually to fit Twitter dataset, and this could justify this decrease. To overcome this limitation, our ongoing research focuses on replacing the gazetteer with a dynamic feature that could be calculated based on the given training set, and consequently, making the network end-to-end, without any dependency of external resources of knowledge.

[3]https://www.cs.cmu.edu/~ark/TweetNLP/

IV. Conclusions and Future Work

In this paper, we demonstrated the effectiveness of a NER system in detecting textual entities in the Tor onion domains. Also, we pointed out the drawback of the state of the art model which allowed us to focus our ongoing research on making the model as an end-to-end network. We found that even if we train the NER model with a dataset that is not related to the final application, it is still capable of detecting useful entities that are related to suspicious activities. In addition to introducing a new dynamic feature that replaces the gazetteer, we plan to build a customized NER model for Tor domains that might help the authorities in monitoring and analyzing the Tor Darknet content. Hereafter, those automatically recognized entities can serve as an input for our previous work in [1] to build a fully-automatic tool for detecting the emerging products in the Tor Darknet.

Acknowledgement

This research is supported by the framework agreement between the University of León and INCIBE (Spanish National Cybersecurity Institute) under Addendum 01. We acknowledge NVIDIA Corporation with the donation of the Titan XP GPU used for this research.

References

[1] Al-Nabki, M., Fidalgo, E., Alegre, E., and Gonzalez-Castro, V., "Detecting Emerging Products in Tor Network Based on K-Shell Graph Decomposition", *III Jornadas Nacionales de Investigación en Ciberseguridad (JNIC)*, vol. 1, no. 1, pp. 24-30, 2017.

[2] Al-Nabki, M., Fidalgo, E., Alegre, E., and Fernández-Robles, L. "Torank: Identifying the most influential suspicious domains in the Tor network". *Expert Systems with Applications*, vol. 123, pp.212–226, 2019.

[3] Al-Nabki, M., Fidalgo, E., Alegre, E., and de Paz, I. "Classifying Illegal Activities on Tor Network Based on Web Textual Contents", *European Chapter of the Association for Computational Linguistics, 2017*.

[4] McCallum, A., and Li, W. (2003, May). "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons". *In Proceedings of the seventh conference on Natural language learning at HLT-NAACL. Association for Computational Linguistics*.

[5] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. "Neural architectures for named entity recognition". *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.As-sociation for Computational Linguistics*, 2016, pp. 260–270.

[6] Ma, X., and Hovy, E. (2016). "End-to-end sequence labeling via bi-directional lstm-cnns-crf." *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* pp. 1064-1074

[7] von Däniken, P., and Cieliebak, M. (2017, September). "Transfer learning and sentence level features for named entity recognition on tweets." *In Proceedings of the 3rd Workshop on Noisy User-generated Text* pp. 166-171.

[8] Aguilar, G., Maharjan, S., Monroy, A. P. L., and Solorio, T. (2017). "A Multi-task Approach for Named Entity Recognition in Social Media Data". *In Proceedings of the 3rd Workshop on Noisy User-generated Text*