Extracting Dark Web Threats Using NER

Qurat ul Aain, Murat Can Ganiz
Department of Computer Engineering
Marmara University
Istanbul, Turkey
q.anniee@gmail.com
murat.ganiz@marmara.edu.tr

Abstract — **Cybersecurity is the practice of defending computers, servers, networks and data from malicious attacks. Understanding means of cyber attacks helps organizations defend themselves. Dark Web provides anonymity to its users which helps cyber criminals to organize systematized attacks without getting easily detected. Dark Web is used by criminals for illegal activities, therefore, analyzing the data from the Dark Web can give valuable insights. Natural Language Processing (NLP) is used to analyze text data. A method of NLP is Named Entity Recognition (NER) which is one of the ways of extracting cyber threat information from the data collected from the Dark Web. This can be done using the data crawled from the Dark Web discussion forums that are related to hacking and annotating that data to train deep learning models to analyze new related data. This project will use BiLSTM CRF as the Deep Learning model to analyze annotated benchmark dataset 'Conll 2003' and will also fine tune the BERT NER pretrained model on the Dark Web data to extract entities from new and unseen data. No other work has systematically classified Cybersecurity related named entities from the Dark Web. In this work, 14 total entities were found from the hacking forums such as exploits, organizations, currency, vulnerability scanner, person, location, protocols, password cracker, packer sniffer, attacks, etc.**

Keywords: Cybersecurity, NER, Dark Web, Deep Learning, BiLSTM CRF, BERT NER

I. INTRODUCTION

Deep Web makes up 96% of the World Wide Web and 57% of it contains illegal and criminal activities [4] and is named the 'Dark Web'. TOR (The Onion Router) is used to access the Dark Web as it provides anonymity to its users and to the criminals. Dark Web is almost untraceable and is adopted by criminals to conduct illegal and malicious activities.

There are several rising threats in the Dark Web today. They include human and sex trafficking, pornography, assassination of high ranking politicians, drug transactions, pedophilia, terrorism like ISIS, stolen data and Darknet currency like Bitcoin [4].

The focus of this project will be on the hacking discussion forums in English Language, where hackers exchange information. The hacking forums include tutorials for hacking and are mostly hosted by hacker organizations like The Anonymous, where novice hackers ask for tips and methods to achieve an end goal, usually of hacking a person or an organization. Studying these forums helps understand their methodologies, means and can even help predict the next offensive attack. The goal is to find a way to prepare defensive strategies against such attacks and to have defense mechanisms in place, since the hackers usually target the weakest link in a network and use it as a launch point for other attacks. Any vulnerability is exploited, therefore it becomes essential to understand cybersecurity and have means defensive mechanisms in place. This project will extract cyber attack information from the Dark Web, focusing on the forums dedicated to hacking.

Cyber attack is an assault launched by cyber-criminals against an organization, a computer or a person. A cyber attack can be malicious, it can disable computers, steal data or used breached computer as a launch point for other attacks.

NER (Named Entity Recognition) is an NLP (Natural Language Processing) technique whereby named entities such as words or phrases are identified from a text data using deep learning techniques. Entities such as person, organization, location, etc.

In this study, 14 entities were found from the Dark Web hacking forums.

The rest of the paper is organized as follows. The background and related work are covered in section II, then the approach for detecting named entities is presented in section III. Section IV provides summary of the experimental study, section V provides future work and conclusion and Section VI includes references.

II. BACKGROUND AND RELATED WORK

Various threats exist in the Dark Web, such as drug transaction, terrorism, human and sex trafficking, assassination, pornography, however, cyber attacks, etc. The Deep Web consists of websites, discussion forums, market places, etc.

The focus of this project is going to be the cyber attacks that are analyzed using the hacking discussion forums. The hacking discussion forums are a place where the hackers share information about cyber attacks and create tutorials for others to learn hacking. It gives us valuable information into the vulnerabilities that the hackers usually aim at. Since it is relatively easy for someone to gain access to this information and use it for malicious purposes, learning about cybersecurity is vital.

In the literature, the techniques used to analyze the Dark Web data include multi-class classification using machine learning techniques such as SVM, KNN, Naive Bayes, and also deep learning techniques such as CNN.

However, there doesn't exist a lot of studies that use Named Entity Recognition (NER) to analyze the Dark Web data. Analyzing the literature, shows that the state of the art for analyzing named entities is the Deep Learning model that uses BiLSTM CRF as encoder-decoder, and also fine-tunes pretrained language model embeddings for task-specific applications.

The literature shows that Named Entity Recognition (NER) can be performed using several different methodologies that combine different permutations of encoder-decoder layers.

NER is an important preprocessing step for downstream applications in NLP tasks. It is used in information extraction and several other NLP tasks [2]. NER identifies text belonging to semantic types such as person, location, organization, proteins, DNA, vulnerability scanner, etc. The NER system is usually evaluated by comparing outputs to the human annotations, by identifying false positives, false negatives and true positives, to compute precision, recall and F-score.

Four main streams of techniques are applied in NER:

1. Rule-based approach with hand-crafted rules,
2. Unsupervised learning which is clustering and is based on context similarity,
3. Feature-based supervised learning where feature engineering is critical step and is CRF-based and it considers the context of words,
4. Deep-learning (DL) based approaches which this paper focuses on.

This paper uses taxonomy of Deep Learning based NER which uses distributed representation for input which is based on dense vectors of low dimensions where each dimension represents a latent feature, distributed representation captures semantic and syntactic properties of words. Second part of DL-based NER is context encoders such as CNN, RNN, neural language models and deep transformers. And lastly, tag decoders such as multi-layer perceptron with softmax, Conditional Random Fields (CRF) and RNN. This paper shows that the most commonly used architecture for NER using Deep learning is BiLSTM-CRF, which is also the state-of-the-art currently. Fine-tuning pre-trained language model embeddings is a new paradigm for neural NER. NER performance can also be boosted by using external knowledge, but this is difficult. Transformer encoder is more effective than LSTM when transformer is pre-trained on huge corpora, transformer encoder is also faster than RNN when length of sequence is smaller than dimensionality of the representation, also RNN and Pointer Net decoder are greedy and so are slower. CRF is the most common choice of tag encoder.

Main challenge of NER is the need of having big annotated data in the training step, which is time-consuming and expensive. Quality and consistency of annotations are also major concerns because different trained models in existence currently are not consistent with one another so models trained on one dataset may not work well on another data. Also, nested entities and fine-grained entities (where one entity is assigned multiple types) are common [2].

Systematic Literature Review on Deep web, narrowing to 65 relevant articles is conducted in another study [4]. This Review shows that the Deep web makes up 96% of the World Wide Web and 57% of deep web contains illegal and criminal activities. TOR (The Onion Router) provides anonymity since the information shared is peer-to-peer and not through a centralized computer, also the relay station in TOR conceals activities. This gives criminals the anonymity for conducting illegal activities. The two research questions were extracted and then answered from the data extracted from these papers and analyzed. The data was then synthesized, by first selecting attributes of each research question and then extracting themes from the articles to answer these research questions: What are the rising threats in Dark Web crimes? 8 major crime threats were found such as Human and sex trafficking, around 0.54% of the world population is under modern slavery and Dark Web contributed to this, other threats were child pornography, assassination of police officers and high ranking politicians, drug transactions, pedophilia, terrorism by ISIS, stolen data and Darknet currency like Bitcoin to launder money. The second question was: What types of techniques are applied to locate criminals on the Dark Web? There are 9 criminal detection methods which include Hash Value Analysis, Sock Puppets and Informant Analysis, Network analysis methodologies by classifying network traffic, Marketplace scraping, Monitoring dark web, Honeypot deployment as detection method for Ransomware to deceive the attackers, Tripwire implementation to detect attack and hacking, Anomaly detection method, and Intrusion detection techniques. Finally, four law enforcement methods were found such as using criminal law, using social media, DAPRA (Defense Advanced Projects Research Agency), Bitcoin Flow and MILAT (Mutual Legal Assistance Treaty).

There are various techniques that are used for detecting malicious activities such as Machine learning technique for monitoring and detecting malicious activities, affect analysis technique of the extremist groups, fingerprinting technique of the hidden service addresses to gather unique addresses, authorship analysis technique for user identification in the drug trafficking area of Dark Web, Social network analysis technique using text mining in terrorism activities, Identity deception detection technique in social media environment using publicly available data on Wikipedia, and Dark web scraping technique to find the lists of the Onion URLs using Reddit and DeepDotWeb websites, using a web crawler written in Apple script.

An operational system is created in another paper for cyber threat intelligence gathering from Deep Web [5]. this system focuses mainly on hacker forum discussions and marketplaces. This system collects 305 cyber threat warnings each week. The system is created using data mining and machine learning techniques. This operational system consists of 1) human analyst found forums and marketplaces on TOR and search engines populated by malicious hackers, 2) crawler is a program that is used to cross the websites and gather topic-based html documents, 3) parser extracts specific information from marketplaces and forums and this structures information is stored in two relational databases, one for marketplaces and one for forums, 4) classifier based on machine learning technique and an expert-labeled dataset, this classifier is integrated into parser to filter out malicious things. Binary classification with the data sample of products on marketplaces, and forum topics, as being relevant or not. This classification is done in a supervised manner using Naive Bayes, Random Forest, Support Vector Machines, and logistic regression. Semi-supervised is better since it can work with less data than supervised tasks. Grid search is done to find optimal parameters for the learning techniques. Only 13% of marketplace exchanges are malicious and criminal, which need to be identified in both the marketplace and the forums. The data processing steps for this are, 1) text cleaning, 2) misspelling and word variations using stemmers, Bag-of-Words, n-gram model, 3) large feature space using sparse feature matrix in spacy library, 4) preserving title feature context by concatenating title and description before extracting features led to worse classification performance, but getting a feature vector of title and description, then horizontally concatenate these vectors gives better results.

Ten marketplaces were used to train and test the learning model [4]. Experimental setup involves leave-one-marketplace-out cross-validation. 25% of labeled data was used for supervised learning tasks for each marketplace. Then performance is evaluated on precision, recall, and F1 score. In supervised tasks, SVM with linear kernel performed the best. The semi-supervised task using the co-training and linear SVM performed slightly better. They created a social network with both the marketplace and the forums. A connected graph was produced using usernames by vendors and users in each domain, and a subgraph of individuals who are both vendors and users. Visualizing this allows the associations to comprehend the hacker networks. The presence of users on multiple markets and forums follows power law (where relative change in one quantity has proportional relative change in the other quantity). Studying these forums gives insights into the social relationships in these communities, where distribution of information is based on skill level and reputation. Another discovery was that tutorials were the most common way of sharing resources for malicious attacks.

III. APPROACH

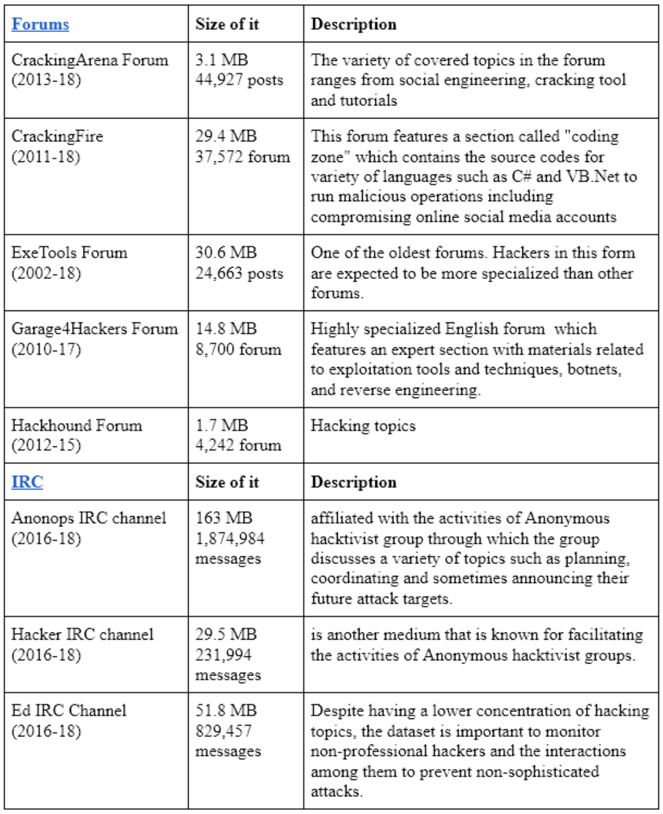
A. Data Collection

The first step of this project was finding relevant data for Named Entity Recognition for Dark Web hacking discussion forums. There doesn't exist a lot of Dark Web data easily available online. Some of the open source datasets were hacked and removed by hacker groups. However, Arizona State University has crawled a number of Dark Web websites and have collected huge amounts of data. This project will use the hacking forums data from the Dark Web, downloaded from the Arizona State University website.

Table 1 shows the summary of the forums and Internet Relay Channels (IRCs) datasets on Arizona State University. The dataset "Hacker IRC" was the main focus of this study.

This open source data is highly useful and relevant to research on the Dark Web. Therefore, for the scope of this study, crawling the Dark Web wasn't necessary. However, the biggest limitation was the size of the forum datasets. They are huge files and since the forums are generally unstructured, they required several preprocessing steps. In order to simplify the process and to create a working framework, only "Hacker IRC channel" was chosen as the main dataset of this project.

Preprocessing this data required a number of NLP preprocessing text cleaning steps such as removing bad lines, removing the names of people who were chatting, and extracting the actual discussion going on. For the scope of this project, the names of people was not essential, also some other information was irrelevant. Finally, this cleaned text data was then further processed using the next steps.



Forums	**Size of it**	**Description**
CrackingArena Forum (2013-18)	3.1 MB 44,927 posts	The variety of covered topics in the forum ranges from social engineering, cracking tool and tutorials
CrackingFire (2011-18)	29.4 MB 37,572 forum	This forum features a section called "coding zone" which contains the source codes for variety of languages such as C# and VB.Net to run malicious operations including compromising online social media accounts
ExeTools Forum (2002-18)	30.6 MB 24,663 posts	One of the oldest forums. Hackers in this form are expected to be more specialized than other forums.
Garage4Hackers Forum (2010-17)	14.8 MB 8,700 forum	Highly specialized English forum which features an expert section with materials related to exploitation tools and techniques, botnets, and reverse engineering.
Hackhound Forum (2012-15)	1.7 MB 4,242 forum	Hacking topics
IRC	**Size of it**	**Description**
Anonops IRC channel (2016-18)	163 MB 1,874,984 messages	affiliated with the activities of Anonymous hacktivist group through which the group discusses a variety of topics such as planning, coordinating and sometimes announcing their future attack targets.
Hacker IRC channel (2016-18)	29.5 MB 231,994 messages	is another medium that is known for facilitating the activities of Anonymous hacktivist groups.
Ed IRC Channel (2016-18)	51.8 MB 829,457 messages	Despite having a lower concentration of hacking topics, the dataset is important to monitor non-professional hackers and the interactions among them to prevent non-sophisticated attacks.

Table 1: Dark Web Datasets

B. Annotating the data

The next step was to convert the unstructured forum discussion format to NER data format. There existed a lot of unseen entities in this dataset that needed to be hand annotated. The entities were not predefined in the literature very extensively. Therefore, this step required thorough understanding and research of different terms that were used, the software and hacking terminology that was being discussed. Finally, categorizing these terms into relevant categories, that were not too broad or too narrow, was the major challenge in this step.

The text data was firstly converted to NER input format, which consisted of three columns; sentence number, words in that sentence in a vertical format, and the named entities in BIO tag format. The input format is in BIO tagged format, where "B" represents beginning, "I" represents inside a tag and "O" represents others. So the word 'Barack Obama' should be tagged as Barack as B-PER and Obama as I-PER.

Table 2 shows the named entities that were extracted, and the words that were labeled with those entities. This step required attention to detail and several trials to conduct systematic and consistent named entities. 14 entities were found in total;

Tags	Meaning	Words labeled using this tag
1. ORG	Organization	Cisco, cryptomills, NSA, FED, Anonymous, Guardian, Yahoo, Vulnhub, Facebook, LE, Google
2. LOC	Location	UK, California, Russia, China, Iran, India
3. PER	Person	Edward Snow, Lauri, DPR (dread pirate roberts), Ahmad Rahami
4. CUR	Currency	Bitcoin
5. SW	Software	SQL, SQLi, CMS, Image- exiftool, teamviewer, Xll-redirect, TOR, TBB, Aircrack-ng, RDP, wireshark, maltego, johntheripper, AdminLoginFinder
6. OS	Operating system	Distro, Kali, Linux, VMS, Windows, Unix, Nexus, VM, Unix
7. FRM	Dark Web Forums	Anonops, silkroad, ircds, ircops, Anonymous, tutorials
8. LNG	Programming Languages	PHP, perl, Python, ASP, ColdFusion
9. SCN	Vulnerability Scanners	Nikto, nmap, uniscan, unix-privesc-checker, wpscan, joomscan, joomla, dork for SQL
10. PRT	Protocol	VNC, VPN, TCP, FTP
11. TOL	Miscellaneous tools	NFAT, cookie, cain and abel, ettercap, MSF, msfconsole, networkmine
12. PTT	Penetration testing tools	Cintruder, nipper, sqlmap, uniscan, xsser, cmsexplorer, dsniff
13. SER	Server	C&C, webproxy, proxy
14. ATK	Attack	Hashcatplus, ddos, Xss, botnet, malware, zeuz, FPD, extrabacon

Table 2: 14 entities from the Dark Web Dataset

At the end, it is common for named entity related datasets to contain highest number of "O" in the BIO tagged data. Figure 2 shows this uneven distribution, and Figure 3 shows the distribution of other entities when "O" was removed.

Figure 4 shows that unlike the Conll-2003 dataset, which and some other academic datasets, this dataset did not have consistent length of sentences.

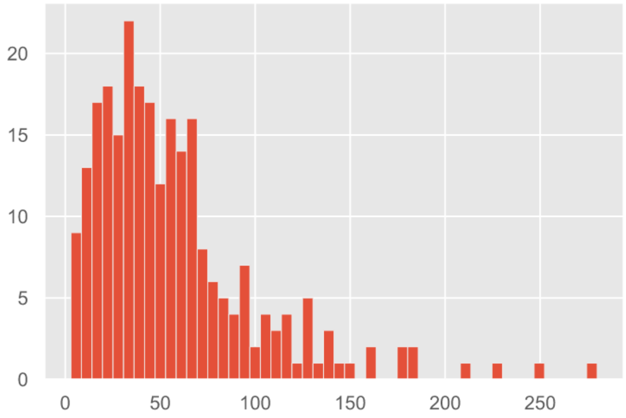


Figure 1: Distribution of length of sentences in the dataset

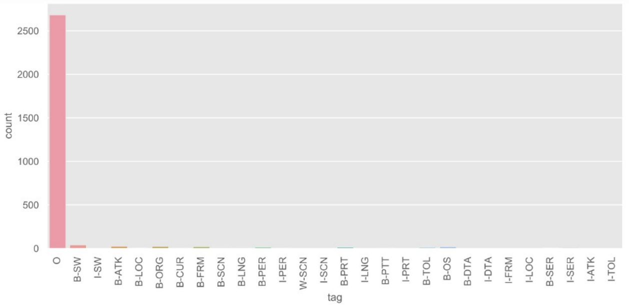


Figure 2: Count of tags including "O"

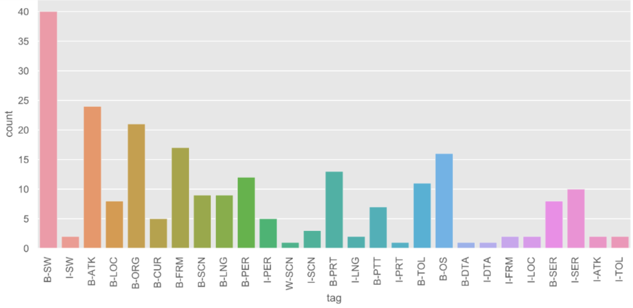


Figure 3: Distribution of tags excluding "O"

C. *BiLSTM CRF for Benchmark Conll-2003 dataset*

The final step was to train the Deep Learning model using this annotated dataset. However, since the entities were hand labeled, the final annotated data was very small in size (35 KB) and therefore the BiLSTM CRF model was not effective. Since BiLSTM CRF model requires data to be divided into three separate datasets namely train, testa and testb.

BiLSTM CRF is the state-of-the-art model currently for the NER task. Therefore, it was used to train on the CONLL 2003 dataset, to set the benchmark results. The model was trained on CPU, it took 24 hours to train and it achieved 98% accuracy, 86% precision and 78% recall. The CONLL 2003 dataset contains just Person, Location and Organization as entities. Figure 4 shows the architecture of this model. The GloVe embeddings were used for word level representation, these along with the words, POS tags and named entities were fed into the encoder which is Bidirectional Long-Short Term memory transformer here. Finally, the decoder was CRF or Conditional Random Fields. CRF is a relational learning model, it calculates conditional probabilities to output named entities.

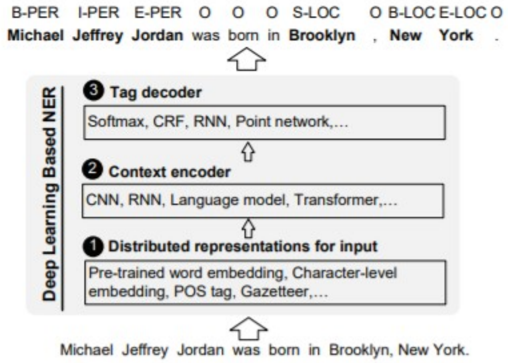

B-PER I-PER E-PER O O O S-LOC O B-LOC E-LOC O
Michael Jeffrey Jordan was born in Brooklyn , New York .
Deep Learning Based NER
3 Tag decoder
Softmax, CRF, RNN, Point network,...
2 Context encoder
CNN, RNN, Language model, Transformer,...
1 Distributed representations for input
Pre-trained word embedding, Character-level embedding, POS tag, Gazetteer,...
Michael Jeffrey Jordan was born in Brooklyn, New York.

Figure 4: BiLSTM CRF model architecture using the Deep Learning taxonomy [2]

Figure 5 shows the formula that CRF uses to calculate conditional probabilities.

Figure 5: CRF uses this formula to calculate conditional probabilities

D. *BERT NER for Dark Web Data*

Since the amount of hand-labeled Dark Web data is limited, BERT NER model is used to train on the Dark Web data. The pretrained model is fine tuned to learn the new entities, and then new sentences are fed into the pipeline and finally, the entities are displayed as outputs.

Using HuggingFace Library and the transformers, BERT Base Uncased (12-layer, 768-hidden, 12-heads, 110M parameters) was implemented, firstly with 10% test and 90% train data split, the validation accuracy is 98% and validation loss is 26%. Looking at the learning curve below, the model is suffering from both high variance and is underfitting the data. The cause for this is that the training data is too small. However, this can be fixed by changing the train-test split.

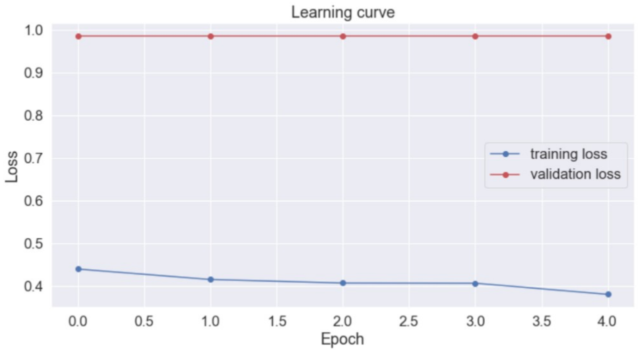


Figure 6: Bert-base-uncased with 90-10 split (underfit)

In “Good-Fit” Learning curves, the loss is lower on the training data than on the validation data and there should be some gap between the two curves. This is called the “generalization gap”. However, continued training of good fit can result in over-fitting.

Using BERT Base Uncased, with 20% test and 80% train data split, the validation accuracy is 98%, validation and loss is 32%. Looking at the learning curve in Figure 7, the model is performing better with this split. However this is not the optimal fit either since the training and validation loss are not close to each other.

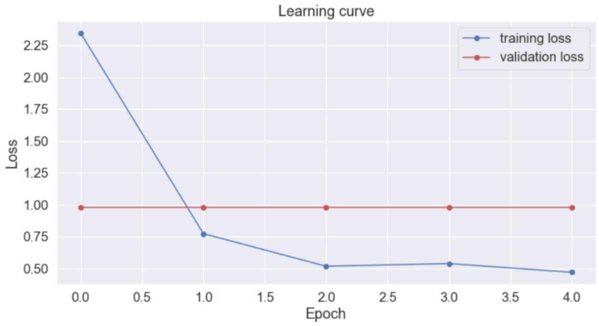


Figure 7: Bert-base-uncased with 80-20 split

Changing the train-test split to 30% test and 70% train data split, the validation accuracy is 97% and validation loss is 46% as shown in Figure 8. Therefore, 20-80 split worked best with this data.

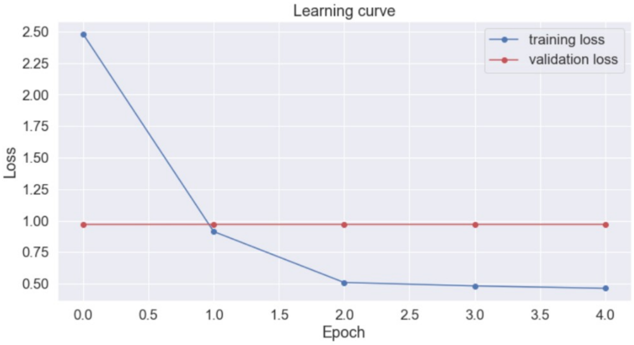


Figure 7: Bert-base-uncased with 70-30 split

Using BERT-base-cased (12-layer, 768-hidden, 12-heads, 109M parameters, trained on cased English text), with 30% test and 70% train data split, the validation accuracy is 97% and validation loss is 44%. Looking at the learning curve in Figure 8, the model is performing better with this split.

In conclusion of these experiments, it was found that BERT-base-uncased with 80-20 split performed best on this dataset.

Finally, when the new sentences were processed and passed through BERT NER model, the entities could be extracted.

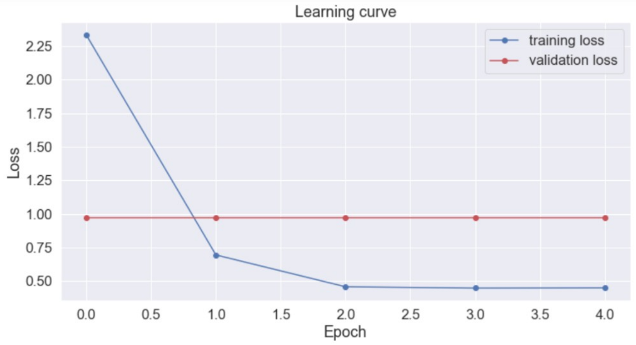


Figure 8: Bert-base-cased with 70-30 split

IV. FUTURE WORK

Future work on this project should include a much larger amount of Dark Web annotated data, something that requires a lot of human error and is also time-consuming.

Moreover, future research in finding other ways of conducting named entity recognition, other than using data in the standard Conll data format, could be helpful.

Also, find-tuning pretrained language model embeddings for neural NER is a new paradigm. Find-tuning the pretrained language model on all of the Dark Web hacking related discussion forums from Arizona State University and releasing it for public use will be helpful.

Future research can also consider fine-grained NER and boundary detection, joint NER and entity linking, deep-learning NER on informal text with auxiliary text, scalability of DL-based NER, deep transfer learning for NER, and an easy-to-use toolkit for DL-based NER.

Forums in different languages, such as in Russian, French, etc., should also be analyzed for entities and hacking methodologies. Currently, this is a field which is ignored especially in context of Dark Web data, and it might include important details that are not seen in the English discussion forums.

Most of the data on the Arizona State University website related to the Dark Web is not up-to-date. So crawling more data from the Dark Web is another area yet to be explored.

Making a pipeline, for example using ElasticSearch, where new data can be preprocessed and named entities can be the output in real-time will be a good way to recognize the potential of named entity recognition.

V. CONCLUSION

Natural language processing including named entity recognition is a relatively new field and new breakthroughs are happening each month. Therefore, there is a lot of potential for exploration and a lot of room for growth in this area. While the current methods perform with almost 98% accuracy, the data preprocessing part is the most challenging area in named entity recognition currently. Another challenge is finding relevant embeddings for a specific field of research, since training embeddings is very costly in terms of time and available GPUs. This project hand annotated Dark Web data to find named entities in Dark Web cyber-attack related discussion forums and found 14 entities. This project also fine-turned the preprinted BERT NER model and found an accuracy of 97% on Dark Web data.

VI. REFERENCES

[1] Ebrahimi, M., Samtani, S., Chai, Y., & Chen, H. (2020, June). Detecting Cyber Threats in Non-English Hacker Forums: An Adversarial Cross-Lingual Knowledge Transfer Approach. IEEE Symposium on Security and Privacy (IEEE S&P), Deep Learning and Security Workshop (DLS), 3(0), 1-7. https://mohammadrezaebrahimi.github.io/publications/DetectingCyberThreatsinNonEnglishHackerForums_AnAdversarialCrossLingualKnowledgeTransferApproach_Ebrahimi.pdf

[2] Li, J., Sun, A., Han, J., & Li, C. (2020, March 18). A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering, 0(0), 1-20. https://arxiv.org/pdf/1812.09449.pdf

[3]Lin, B. Y., Xu, F. F., Luo, Z., & Zhu, K. Q. (2017). Multi-channel BiLSTM-CRF Model forEmerging Named Entity Recognition in Social Media. *Shanghai Jiao Tong University*, *0*(0), 1-6. https://www.aclweb.org/anthology/W17-4421.pdf

[4] Nazah, S., Huda, S., Abawajy, J., & Mehedi Hassan, M. (2020, Sept 15). Evolution of Dark Web Threat Analysis and Detection: A Systematic Approach. IEEE Access, 8(0), 171796-171819. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9197590

[5] Nunes, E., Diab, A., Gunn, A., Marin, E., Mishra, V., Paliath, V., Robertson, J., Shakarian, J., Thart, A., & Shakarian, P. (2016). Darknet and Deepnet Mining for Proactive Cybersecurity Threat Intelligence. IEEE Conference on Intelligence and Security Informatics (ISI), 0(0), 7-12. https://arxiv.org/pdf/1607.08583.pdf

[6] Pastor Galindo, J., Nespoli, P., Gomez Marmol, F., & Martinez Perez, G. (2020, Jan 9). The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends. IEEE Access, 8(0), 10282-10304. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8954668

[7] Schäfer, M., Fuchs, M., Strohmeier, M., Engel, M., Liechti, M., & Lenders, V. (2019, May). BlackWidow: Monitoring the Dark Web for Cyber Security Information. 2019 11th International Conference on Cyber Conflict: Silent Battle, 900(11), 1-21. https://ccdcoe.org/uploads/2019/06/ART_27_BlackWidow.pdf