A Systems Approach to Cyber Security
A. Roychoudhury and Y. Liu (Eds.)

doi:10.3233/978-1-61499-744-3-57

A Deep Convolutional Neural Network for Anomalous Online Forum Incident Classification

Victor POMPONIU [1] and Vrizlynn L. L. THING
Cyber Security and Intelligence (CSI) Unit
Institute for Infocomm Research (I^2R)
*Agency of Science Technology and Research (A*STAR), Singapore*

Abstract. Web forums are a frequent way of sharing useful information among people. They are becoming the main source of up-to-date information and marketplaces pertaining to different domains, including criminal content and zero-day security exploits. Analyzing the web forums of the existing discussion threads is an alternative method to understand the exploits and fraud modalities a law breaker will most likely make use and how to defend against them. However, in many cases, it is hard to capture all the relevant context of the forums which is needed for classification. In this paper, we introduce a data-driven technique to mine the web forums and provide policy recommendations to the defender. A neural network (NN) is used to learn the set of features for forum classification. Furthermore, we present the evaluation and results from employing our method, with various system configurations, on real-world datasets collected form the web.

Keywords. analytics, classification, forum, cyber security intelligence, neural networks, word embedding.

1. Introduction

The Internet has become the place where the users research, purchase, socialize and learn about the world by a few clicks of a mouse or keystrokes. In this cyberspace users, leave behind rich trails of behavioral information [6] and activity intent, which can be mined by third-party trackers such as law enforcement agencies, social networking websites, and data analytic providers.

However, the web has also a less known secretive component [4] where dark users can socialize, access and share concealed services [5]. Regardless of how the information is expressed, can we extract deep insights form this alternate network? What are the most efficient techniques to analyze criminal activities, the modalities of committing them and discover who make use of this hidden networks? This part of the Internet started to enter in the spotlight of the public due to the media that raised awareness of several cases such

[1]Corresponding Author: Victor Pomponiu, Cyber Security and Intelligence Unit, Institute for Infocomm Research, Agency of Science Technology and Research, 1 Fusionopolis Way, Singapore 138632, Singapore; E-mail: v-pomponiu@i2r.a-star.edu.sg.

as releasing of NSA hacking tools [1], card cloning services [24] and online illegal drugs selling stores [2].

The most common method to systematically extract data from the websites is to use an automated program, called crawler [7]. It requires, for each website, custom parameter definitions of the elements of interest and a database to store the crawled data. The aim of security intelligence analytics is to proactively detect the pattern of the fraud activities by processing the vast amount of data (text, image. etc.) collected by the crawler.

Social network analysis [30] and machine learning are the most used tools to mine web data, with a tremendous impact on our everyday life. This new direction infused by artificial intelligence (AI) is driven by a paradigm shift in system design. Instead of learning hand-crafted features, which involves deep domain knowledge, the current approach is able to capture hierarchical feature representation automatically extracted from vast amounts of example data which leads to significant performance improvement in fields likes speech understanding, computer vision, and human language processing.

In this paper, we are investigating the feasibility of AI for security intelligence analysis of the web forums. In particular, a neural network is trained with the phrases of the post for forum classification. The classification performance of the method attains good accuracy and shows that it is able to understand the fraudulent forum posts created by manipulators.

Our contribution to the field are twofold and summarized below:

- We propose a methodology that employs a neural network to learn deep features from the forums threads which can be later used for security intelligence analysis. Furthermore, we investigate the suitability of adapting the knowledge of the models pretrained on different domains to the peculiar the domain of illicit content detection.
- Extensive experiments are carried out to illustrate that the proposed method outperforms the state-of-the-art baselines algorithms.

The remaining sections of the paper are structured as follows. Section 2 reviews web forum analysis methods, with a focus on those that are targeting those with fraudulent content. In section 3 we formalize the problem that we are addressing, while section 4 presents in detail our proposed system. Section 5 presents the experiments and results obtained. Finally, Section 6 discusses the challenges of the system and concludes the paper by highlighting several future research directions.

2. Related Works

The extraction and analysis of the web discussion forums has acquired a lot of attention and currently is an active research field. Nowadays, the available solutions share many characteristics and the only differences are the level of automation, the type of pattern recognition (whether is classification or retrieval) and the domain from where the data is collected. In this section, we give a brief overview of the recent state-of-the-art methods, while paying a particular attention to those that are devoted to detect illicit activities patterns in web forums.

The problem of searching similar threads to a given thread, have been first addressed by Singh et al. [9]. Their framework represents the thread structure via a graph model,

built from forum posts, to which they incorporate heuristics that can capture the thread similarity. It is worthwhile to point out that, the model is able to cope with the issues that arise when analyzing the forums such as the drift of the post's subject and their dimensionality within the threads.

Some of the earliest works [10] on web forum classification tried to adapt the Latent Dirichlet Allocation (LDA) approach for modeling the topic of the threads. The topic distribution inferred from the contextual data was later used as a feature descriptor for thread classification.

[11] introduces a method that analyzes the forum posts [12] in order to detect patterns of terrorist activities. At the core of the method is a hybridized feature selection algorithm which takes into account the standard features used for text mining such as term frequency (TF), document frequency (DF), term frequency-inverse document frequency (TF-IDF) and entropy. After extracting the feature sets, the scheme employs a feature selection algorithm based on the union combination and symmetric difference functions. To reduce the dimensionality, these functions weighs the features according to a hybrid criterion. The experimental results performed on the Web Forum dataset confirms the ability of the proposed feature selection to identify a reduced set of discriminant features that can be used for forum classification.

Based on the naive Bayes (NB) classifier, in [13], another feature weighting approach was proposed. In a nutshell, the idea is to integrate the feature weights learned from the training data as prior information to aid in the estimation of the conditional probabilities of naive Bayes. The experimental simulations carried out on several datasets from the UCI repository show that the modified NB improves the performance when compared to other state-of-the-art NB text classifiers. However, the NB classifier is quite rigid since it assumes linearity of the model and statistical independence of the features.

Diab et al. [14] devised a system that collects information used to early identify threats from the Internet websites such, forum discussion boards and marketplaces that sell illegal goods. The system focuses more on the integration and deployment issues rather than on novel content features for forum classification. The main classification features are bag-of-words and the n-grams extracted from both the title and description of the posts (concatenated in a single feature vector). A preprocessing operation is employed to remove non-alphabetic characters and misspelled words. To cope with the insufficient labeled samples, which are difficult to obtain, the method merges supervised training with semi-supervised techniques, such as label propagation [15] and co-training[16]. During the evaluation, the model was trained and tested on 10 marketplaces, using 3 types of classifiers, i.e., NB, logistic regression and SVM.

In [17], a system targeting the deep dark web forums was devised. The paper analyzed the forums from a network perspective, gathering temporal data and further generating off-line analytics of the social network communities. In a subsequent work [18], they adopted a systematic approach to test the data mining techniques suitable for these forums.

[19] following a game theoretic approach, shows that an attack can be anticipated and, in case it occurs, the damages minimized if the cyber defense system incorporates updated information from the illicit Web marketplaces. The main assumption is that an adversary will search through the available tools and vulnerabilities in the marketplace, and therefore will initiate attacks by leveraging this knowledge. Thus, the security turns into a game, where the defender tries to predict the possible attacks by analyzing the

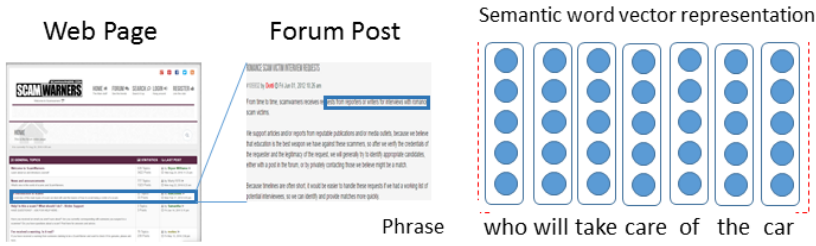

Figure 1. The main components of a web forum page: threads and posts. We represent the post content by a series of sentences and their words.

same base knowledge used by the attacker. In the same spirit but from a socio-economical perspective, other works [20,21] studied the 'gadgets' available in the hackers forums and their implied associated risk.

Other studies [22,23,24,25,26,27,28] chose a completely different approach to tackle the problem, by modeling the online communities (i.e., a social network) as a network graph [29] where each node represents a user and edges are the connections among them, weighted by a similarity feature. Using tools from social network analysis [30], these methods are able to extract, without supervision, insights about the dynamics of the social relationships, the information sharing rate, virality, the illicit content (such as cards, hacking materials etc.), user importance (centrality) within the network and the trust relationships emerged between mutually parties.

Law enforcement agencies are also leveraging on analyzing social networks at large scale to identify, track and counter criminal activities such as money laundering and human trafficking (e.g., the MEMEX program [3] run by DARPA).

3. Problem Statement

A website forum **R** consists of a set of discussion threads **T**, and in each thread, there are several posts **P**. We define a post as the smallest piece of communication generated by a user in online social networks. An illustration of the thread and post of the website is shown in **Figure 1**. Each post has embedded metadata like date and time of posting, the owner of the post etc.

Generally a thread commences with a main post (i.e., the entry post), whose title is associated with the thread title, and comprises all posts that were placed in reply to the main post. It is worth to point out that, the posts in reply to a post contains all the posts already in the thread post. For a forum thread, a graph can be constructed, in which the vertices represent the posts in the thread and edges the links between a post and all responses to it.

The goal of the study is to devise a system that identifies to which set of predefined classes **K** the posts belong. Given a set of **P** and their associated class labels l, the system construct a classifier which predicts the label of an unseen post P_i as follows:

$$\hat{l} = \text{argmax}_k F_k(P_i) \tag{1}$$

where $\hat{l}$ denotes the predicted label, F_k is the classifier and $k \in 1, ..., K$.

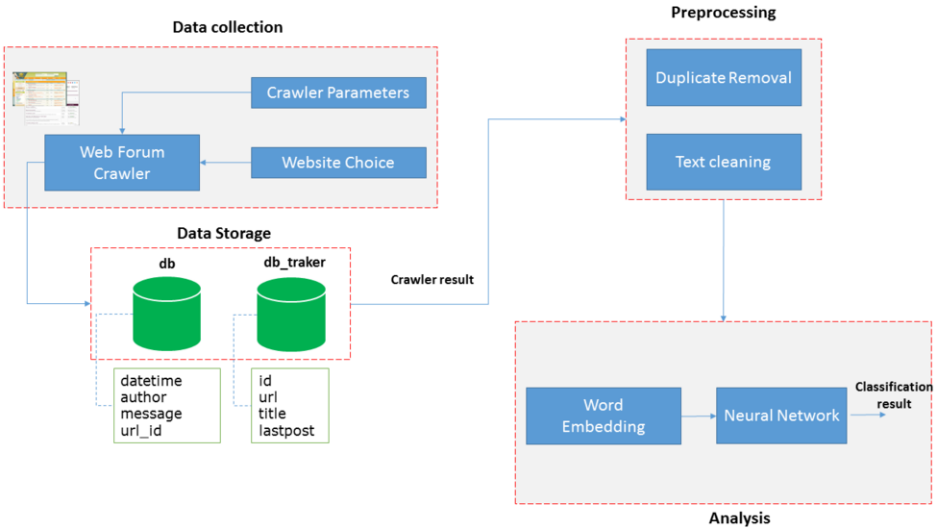

Figure 2. The flowchart of the web crawler system which incorporates the proposed analysis method using a neural network module.

4. Proposed Method

It is challenging to achieve accurate classification of web forums due to their inherent unstructured data. Instead of using hand-crafted features, we propose a modular approach to classify the forums with a combination of thread representation and a deep convolutional neural network.

More precisely, with the aid of custom web crawlers, we collect relevant information from a selection of websites and accumulate the resulting data in structured databases. Then, to simplify the processing and remove any noise from the data we, employ a preprocessing step to the crawled data. The filtered output is transformed to another representation and fed into a neural network for forum classification task.

An illustration of entire the system is depicted in **Figure 2**. In the next sections, we provide a detailed description of the main modules of the proposed system.

4.1. Data Collection

A crawler is a piece of software used to visit and retrieve websites, being a desirable tool which facilitates data collection. In order to initiate a targeted data extraction campaign from the web, we need to specify for the crawler:

1. The pool of websites from where we are interested to collect data. For each website, we do not need to define the web pages since the crawler is able to find and traverse them by using as the starting point the initial website.
2. A set of parameters which defines the elements that need to be localized and retrieved from the selected websites. It is worth pointing out that, this set of parameters are custom for each website, since the structure and context of the elements vary among the websites.

3. For each website, we maintain two relational databases (*db* and *db_tracker*) to store the information such as data and time of the post, author, the entire post message, the url from where the post was retrieved, the title of the thread, and other log data.

4.2. Preprocessing

Usually, thread title and the post descriptions on web forums are flooded with unrelated text and graphics which may perturb the analysis, since they are acting as noise. To tackle this issue, we apply a text cleaning procedure to remove all the non-alphanumeric characters from the title and post. In addition, after this step, we removed all the words that are less then two characters in length.

Another problem that we encounter in processing the thread and posts was the duplicates that frequently occur on forums. To avoid any bias, we decided to remove all the duplicated from data. The last measure that we adopted was to find a suitable representation [31] of the post which is able to cope with the misspellings and word variations.

4.3. Analysis

In our system, each post is described by a set of sentences that are comprised of an ordered list of words, the smallest unit of the post. Each word is transformed to a vector of numbers, called *word vector*. A word vector denotes a word‘s meaning as it refers to other words, by means of a single array of numbers.

To embed the word in vector space representation, a shallow neural network is used which learns the context through recurrent guesses. One example of such a network is word2vec [31] - a two-layer neural net. The input of the network are words and its output is a dictionary in which each word has a vector affiliated to it. The vector which represents the words are called *neural word embeddings*, i.e., a word is mapped to a number. In the vector space generated, close semantical related words have similar vector representations.

The training is done by comparing the words, belonging to the input, against each other. Thus, the target word is predicted using the context (i.e, the *bag-of-words* technique) or another word (i.e., the *n-gram* technique). If a word context can not be predicted by the feature vector, due to reconstitution error, then its components are updated.

At the core of the model that we are using to analyze the forums is a convolutional neural network integrated into the work-flow of the system depicted in **Figure 2**.

Let's consider $s_j \in R^m$ the j^{th}-phrase of a sentence from a forum post and m is the number of words in the phrase. Then, the sentence is represented as:

$$s_j = v_1^j \parallel v_2^j \parallel ... \parallel v_m^j \tag{2}$$

where $v_i^j \in R^n$ is the continuous word vector representation of the i^{th}-word of the phrase s_j, n is the length of the word vector, and $\parallel$ denotes the concatenation operator. This continuous feature representation of the words fits properly with the NN since the non-linear activation of the units is also continuous. We assign the label of the post to all the sentences generated.

We investigated two possible premises for the input of to the NN. In the former setting, suitable for a supervised learning mode, each word vector was randomly initialized

from a Gaussian distribution, i.e., $v_i \sim \mathcal{N}(0, \sigma^2)$, stacked into a matrix and subsequently changed by the network. In the latter initialization approach, we are using a pretrained set of word vectors [31] learned in an unsupervised fashion over a large dataset. These word vectors grasp the semantic and, to a limited extent, syntactic information from their co-occurrence statistics.

A convolution in the word vector space involves employing a filter $w \in R^{n \cdot b}$ on overlapping block of words of size b to generate a new feature representation:

$$\phi_i = f(w \cdot [v_1^j \parallel v_2^j \parallel ... \parallel v_{b-1}^j] + \beta) \tag{3}$$

where ϕ_i is the new generated feature, β is the bias term with real values and f is the non-linearity function such as hyperbolic tangent. After applying the filter on all combinations of overlapping word blocks, we obtain the feature map $\phi = [\phi_1, \phi_2, ..., \phi_{m-b+1}]$. Inspired by the convolutional neural networks applied to object detection, the models enrich the feature representation in the first convolutional layer by generating multiple feature maps using different filters on the input word vectors.

To increase the robustness of the algorithm, we are employing a max pooling operation over the feature map, that is $\hat{\phi} = max(\phi)$, to compute the final feature. Furthermore, in order to capture multiple hierarchical features, we make use of a bank of filters with various parameters, each of the filter generating its own feature map.

A fully connected network layer takes these features and combine them, and output the probability of distribution over the classes. Computing the reconstruction error at this layer will backpropagate and affect both the parameters of the NN and the vector word representations.

When learning using the scam forums, the word embeddings are adapted to characterize the intent and less the syntactics. The hierarchical structure generated by the NN aims to capture most of the information contained in the words, without the need of taking into account the syntactic constraints.

Therefore, the neural networks that we are using, has the following architecture:

1. Input layer, of size m x n, where m is the number of words in the sentence and n is dimension of the word vector.
2. Convolution layer, which filters a window of word vectors with filters of different widths.
3. Max-pooling layer, which captures the relevant features, by taking the maximal value over the feature map.
4. Fully connected layer, which computes the probability distribution over the classes.

In order to prevent overfitting, at the penultimate layer of the network, we employ regularization with dropout by constraining the l_2-norm of the weight vectors. The dropout operation randomly sets to zero (i.e., no weight) a percentage η of hidden units during forward backpropagation step. Thus, for the penultimate layer $\zeta = [\hat{\phi_1}, \hat{\phi_2}, ..., \hat{\phi_m}]$, we modify the output y during the forward phase with

$$y = w \cdot (\zeta \circ \kappa) + \beta \tag{4}$$

where $\circ$ denotes the element-wise multiplication operator and κ is random vector with probability η being 1, which is used to cancel the contribution of the weight vectors.

Table 1. An overview of the datasets used for our experiments. We mostly collected data from English language forum related to security intelligence and general topics.

Dataset	Source	Number of samples	Target classes
Fraud	allscamsforum	1930	7
	scamwarners	2000	
	complaintboard	393	
	ripoffreport	383	
	general forums	5331	
SPAM	spam messages	5574 (with 747 spam)	2

It is worth to notice that, during the training, the network gradients are backpropagated merely via the non-zero units. At inference step, to assess the unseen samples, we will use a scaled version of the learned weight, like $\tilde{w} = \eta \cdot w$, without applying the dropout. Furthermore, after a gradient descent step, we re-scale the weight vector if the following condition is met: $\| w \|_2 = \sigma$ if $\| w \|_2 > \sigma$.

5. Evaluation and Results

5.1. Dataset

The main dataset, used in this study to evaluate the proposed scheme, called **Fraud**, consisting of approximative 13k posts, is obtained by crawling several security related forum websites containing scam, fraud and phishing information. All the webpages have undergone preprocessing and posts found in different webpages but belonging to the same thread were identified. For each thread, the crawler captured information such as the title, first post, other posts, author information and time stamp. The total number of phrases extracted from the **Fraud** dataset was 46784.

In addition, we added to this dataset 5331 forum posts collected from general websites (e.g, movie reviews [33], car forums etc.) without being directly related to any criminal content.

Beside this dataset, we also included a dataset containing a collection of email messages [32] classified into two categories. The main choice of including this dataset was driven by the assumption that there is a relationship between the scam and spam messages, since often scam messages are disguised via spam campaigns. More broadly, we are interested to explore whether the learned patterns (i.e., the semantical words vectors and the syntactics) of the model will enable to detect the illicit (or undesired) character of a sentence, paragraph or a document.

A detailed description of the datasets used for the evaluation and experiments is shown in **Table 1**.

5.2. Selection of the Parameters

For all the simulations we use the following parameters:

1. *General parameters*. We split each sentence of the post in a block of maximum length of 55 words, and the size of the word vector was set to 300. For classification

Table 2. Classification accuracy of the proposed system on the considered dataset. The $Fraud_6$ dataset considers only the post pertaining to scam forums (i.e., six classes) while $Fraud_7$ includes also the general posts. The values in bold correspond to the best performance.

Model	$Fraud_6$	$Fraud_7$
CNN-R1	0.7720	0.8008
CNN-WV1	0.7951	0.8213
CNN-WV2	**0.8073**	**0.8341**

we use 4-folds cross-validation, i.e., we split entire dataset in 4 folds that are used for training and the remaining data is used for testing the model.

2. *NN parameters*. For the neural network, we set the weight decay to 0.95, the number of epochs to 15, the units used were rectified linear (i.e., ReLU), the batch size to 50, the dropout rate to 0.5 and hidden units to 100. In addition, we use filters of size $\{3,4,5\}$ during the convolution and we rescale the l_2-norm of the weights by 9. These parameters values were selected through a grid search procedure on a subset from the **Fraud** dataset. The training of the NN is done through stochastic gradient descent using shuffled mini-batches with the Adadelta update optimization rule [34].

We classify the content of the **Fraud** dataset in 8 classes: *complaints* (posts in these class are related to complains issues), *phishing* (posts in these class are related to phishing attempts), *business scam*, *e-scam* (posts in these class are related to scams that originate from fraudulent emails), *phone-scam* (posts in these class are related to scams that originate from fraudulent phone calls), *prevention* (posts in these class are related to prevention) and *general* (posts that are related to general topics). Instead, for the **SPAM** dataset, we are using just two classes (binary classification): spam and not-spam.

In order to evaluate the classification results, we use several standard measures such as sensitivity (i.e., Sens = TP/(TP + FN)), specificity (i.e., Spec = FN/(FN + FP)) and accuracy (i.e., Acc = (Sens + Spec)/2).

5.3. Classification results

We tried different set-ups for the NN, especially by changing the input representation and the internal parameters of the net. For instance, in two of the configurations we use pretrained features vectors to initialize the neural network. The reason of adopting this approach is due to the lack of labeled data during training, and the assumption that the pretrained vectors have learned a representation that can be transferred to similar tasks.

The pretrained vectors were generated using the word2vec model trained over the Google News dataset [31]. The model has feature vector dimensionality of 300 and uses the bag-of-words technique for training. In case a word vector is not covered by the set of pretrained vectors, it will be randomly initialized with the same variance as the pretrained ones.

Therefore, the model configurations that were tested are:

- *CNN-R*1. This is the baseline method which has all the word vectors randomly initialized, and updated during the training. In this scenario, the convolutional NN model is used both for training (learning the feature vectors from the dataset) and testing (inference).

Table 3. Classification result of the proposed system against all other evaluated classifiers on the considered **SPAM** dataset. The values in bold correspond to the best performance.

Model	**SPAM**
EM	0.8554
MDL	0.9626
Boosted NB	0.9750
Linear SVM	0.9764
Linear SVM-P	0.9582
CNN-R1	0.9573
CNN-WV1	0.9769
CNN-WV2	**0.9822**

- *CNN-WV*1. The model adopted uses the pretrained vectors, and merely updates the parameters of the NN during inference.
- *CNN-WV*2. In this setting, the model starts with the set of pretrained vectors, but fine-tunes the pretrained vectors to better adapt them to the specific domain.

To estimate the generalization error of the classification models based on the selected features, we employed the 5-fold cross validation technique on the set of samples in the datasets. For 5-fold, in each trial, 4 folds are held out and used for training. The remaining fold is used for testing.

The classification results of the evaluated models are shown in **Table 2** on the **Fraud** dataset for different number of classes. The model *CNN-R*1 which is characterized by having all feature words randomly initialized has satisfactory accuracy. However, by using pretrained vectors, we achieved important performance improvements on all datasets.

For the **SPAM** dataset evaluation, which is a binary classification problem, we compared our model with well-known machine learning algorithms such linear SVM, SVM with pretrained word vectors (called linear SVM-P), Minimum Description Length (MDL), boosted NB and Expectation-Minimization (EM). Since this is an imbalanced dataset we also included among the analyzed methods the trivial rejector, as a baseline, applied to the spam class. The comparison results for the NN models against all other evaluated classifiers are presented in **Table 3**.

Although we design a simple and fast architecture, with any sophisticated pooling techniques, the model shows promising results and potential to scale to very large datasets. Overall, the experiments prove that the pretrained vectors are valuable features which can aid in better classification over the datasets. It is worth pointing out that, better performance can be obtained by adapting the pretrained vectors for the task at hand, like is done by the model *CNN-WV*2.

5.4. Discussion

The proposed method is generating word vector representation which is able to infer the contexts in which the words occur. To adapt to a specific NLP task, the neural word model fine tunes these vectors via supervised learning.

The way we represent the feature word vectors plays an important role in the overall model design. Our option for the pretrained vectors was those generated by the word2vec model trained on the Google News dataset [31].

We stress out that this model, which preserves the word order, has a limited ability is measure the interaction between the input vectors, and is weak at understanding the meaning of longer phrases.

It will be interesting to investigate also with other set of pretrained vectors trained on datasets different from the Google News dataset, which can properly acquire more complex linguistic manifestations. Furthermore, another important aspect which may affect the training is the way we initialize the words vectors which are not found within the word2vec model, by taking into consideration the distribution of pretrained vectors.

For the gradient optimization technique, we adopted the Adadelta rule, which is less offensive in the gradient update step than Adagrad, but is faster since it requires less epochs.

In addition, the employed dropout in the network layer is a good option for regularization. Finally, the NN is able to accommodate more features due to the use of a filter with various widths which generate multiple feature maps.

6. Conclusions and Future Work

In this paper, we study the problem of classifying the discussion forum threads related to fraud and other criminal activities into different classes for security intelligence gathering and analysis. Forum post classification has numerous applications such as enabling security experts to quickly filter the elements of interest from the clutter that abound forum threads, to filter harmful contents, to detect cyber-threats and discover emergent exploitation capabilities.

The core of our forum analysis method resolves around a convolutional neural network, which was tested in different configurations, and using several types of word feature vectors. Through a series of experiments, on real world datasets we prove the efficiency of our technique. These results confirms the general idea that using pretrain word vectors is an efficient asset in deep learning for various text classification tasks.

There are several research directions that we are investigating. For instance, adapt the model to extract complex linguistic phenomena from the scam forum. This weak signal could represent the well-crafted illicit intent of exploiting an innocent person, disguisedly into a forum post. To achieve this level of intelligence the model needs to go beyond the word-level, by learning vector representations of paragraphs and sentences, and their hierarchical structure.

We believe that the complexity of the illicit intent, which manifests in the forum posts, can not be full characterized by one dimensional scale. Fine-grained classification of the scams forum can help to better segregate the compositional semantic effects used by this criminals to deceive innocent people.

In the current proposal, we neglect the multimedia content that is attached to the forum threads. We believe that more rich insights can be identified by a multi-modality approach that integrates in a holistic manner both the text and the multimedia content of the forum post.

Acknowledgment

This material is based on research work supported by the Singapore National Research Foundation (NRF) under NCR Award No. NRF-2014NCR-NCR001-034.

References

[1] https://www.washingtonpost.com/world/national-security/powerful-nsa-hacking-tools-have-been-revealed-online/2016/08/16/bce4f974-63c7-11e6-96c0-37533479f3f5_story.html
[2] D. Bradbury. Unveiling the dark web. *Network Security* **2014(4)**, 14-17, 2014.
[3] MEMEX program. http://opencatalog.darpa.mil/MEMEX.html
[4] Threats Report *McAfee Lab Technical Report*. Available at http://www.mcafee.com/hk/resources/reports/rp-quarterly-threats-may-2016.pdf. Retrieved on August 23, 2016.
[5] T. Jordan and P. Taylor. *A sociology of hackers*. The Sociological Review, **46(4)**, 757-780, 1998.
[6] F. Roesner, T. Kohno, and D. Wetherall. Detecting and defending against third-party tracking on the web. *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (NSDI'12)*, Berkeley, CA, USA, 12-12, 2012.
[7] F. Menczer, G. Pant, and P. Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions on Internet Technology (TOIT)*, **4(4)**, 378-419, 2004.
[8] J. Healey. Winning and Losing in Cyberspace, *In Proceedings of the 8th International Conference on Cyber Conflict: Cyber Power*, NATO, pages 37-51, 2016.
[9] A. Singh, P. Deepak, and D. Raghu. Retrieving similar discussion forum threads: a structure based approach. *In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12)*, pages 135-144, 2012.
[10] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to Classify Short and Sparse Text and Web with Hidden Topics from Large-scale Data Collections. *In Proceedings of WWW Conference*, 2008.
[11] T. Sabbah, A. Selamat, Md. H. Selamat, R. Ibrahim, and H. Fujita. Hybridized term-weighting method for Dark Web classification. *Neurocomputing* **173**, 1908-1926, 2016.
[12] Y. Zhang, S. Zeng, L. Fan, Y. Dang, C. A. Larson, and H. Chen. Dark web forums portal: searching and analyzing Jihadist forums. *In Proceedings of the 2009 IEEE international conference on Intelligence and security informatics (ISI'09)*, pages 71-76, 2009.
[13] L. Jiang, C. Li, S. Wang, and L. Zhang. Deep feature weighting for naive Bayes and its application to text classification. *Engineering Applications of Artificial Intelligence* **52**, 26-39, 2016.
[14] E. Diab, A. Gunn, A. Marin, E. Mishra, V. Paliath, V. Robertson, J. Shakarian, J. Thart, A. Shakarian. Darknet, Deepnet Mining for Proactive Cybersecurity Threat Intelligence. *ArXiv e-prints*, 2016.
[15] H. Cheng, and Z. Liu. Sparsity induced similarity measure for label propagation. *In Proceedings of ICCV*, pages 317-324, 2009.
[16] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. *In Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 92-100, 1998.
[17] T. Fu, A. Abbasi, and H. Chen. A focused crawler for dark web forums. *Journal of the American Society for Information Science and Technology* **61(6)**, 1213-1231, 2010.
[18] H. Chen. *Dark web: Exploring and data mining the dark side of the web*. Springer Science and Business Media, volume 30, 2011.
[19] J. Robertson, V. Paliath, J. Shakarian, A. Thart, and P. Shakarian. Data driven game theoretic cyber threat mitigation. *In Proceedings of Innovative Applications of Artificial Intelligence Conference*, pages 4041-4046, 2016.
[20] S. Samtani, R. Chinn, and H. Chen. Exploring hacker assets in underground forums. *In Proceedings of Intelligence and Security Informatics Conference*, pages 31-36, 2015.
[21] E. Marin, A. Diab, and P. Shakarian. Product Offerings in Malicious Hacker Markets. *ArXiv e-prints*, arXiv:1607.07903, 2016.
[22] E Ferrara, W-Q Wang, O Varol, A Flammini, and A Galstyan. Predicting online extremism, content adopters, and interaction reciprocity. *SocInfo 2016: 8th International Conference on Social Informatics*, 2016

[23] C. Fachkha. Security Monitoring of the Cyber Space. *ArXiv e-prints*, arXiv:1608.01468, 2016.

[24] A. Haslebacher, J. Onaolapo, and G. Stringhini. All Your Cards Are Belong To Us: Understanding Online Carding Forums. *ArXiv e-prints*, arXiv:1607.00117, 2016.

[25] L. Le, E. Ferrara, and A. Flammini. On Predictability of Rare Events Leveraging Social Media: A Machine Learning Perspective. *In Proceedings of the 2015 ACM on Conference on Online Social Networks (COSN '15)*, pages 3-13, 2015.

[26] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker. An analysis of underground forums. *In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 71-80, 2011. reference

[27] T. J. Holt, D. Strumsky, O. Smirnova, and M. Kilger. Examining the social networks of malware writers and hackers. *International Journal of Cyber Criminology* **6(1)**, 891-903, 2012.

[28] D. Lacey and P. M. Salmon. Its dark in there: Using systems analysis to investigate trust and engagement in dark web forums. *In Engineering Psychology and Cognitive Ergonomics, volume 9174 of Lecture Notes in Computer Science*, 117-128, 2015.

[29] A. Rajaraman, and J. D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA, 2011.

[30] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. *In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*, pages 497-506, 2009.

[31] T. Mikolov, I. Sutskever, and J Dean. Distributed Representations of Words and Phrases and their Compositionality. *In Proceedings of NIPS*, 2013.

[32] T.A. Almeida, J.M. Gmez Hidalgo, and A. Yamakami. Contributions to the study of SMS Spam Filtering: New Collection and Results. *In Proceedings of the 2011 ACM Symposium on Document Engineering (ACM DOCENG'11)*, 2011.

[33] B. Pang, and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *In Proceedings of ACL*, 2005.

[34] M.D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. *ArXiv e-prints*, 1212.5701, 2012.