Deep Learning Based Phishing E-mail Detection
CEN-Deepspam

Hiransha M, Nidhin A Unnithan, Vinayakumar R, Soman KP
Center for Computational Engineering and Networking(CEN),
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
hiransham5600@gmail.com, nidhinkittu5470@gmail.com

Abstract

Email communication, has now become an inevitable communication tool in our daily life. Especially for finance sector, communication through email plays an important role in their businesses. So, it is very important to classify emails based on their behavior. Email phishing one of most dangerous Internet phenomenon that cause various problems to business class mainly to finance sector. This type of emails steals our valuable information without our permission, more over we won't be aware of such an act even if it has been occurred. In this paper, we reveal about how to distinguish phishing emails from legitimate mails. Dataset had two types of email texts one with header and other without header. We used Keras Word Embedding and Convolutional Neural Network to build our model.

1 Introduction

The internet has become an efficient powerful tool to the present world. Considering the uncontrolled growth of internet and abundant use of emails, has increased insecurity in email communication. We are very familiar with the name spamming whenever we are on the topic email. Spamming is nothing but a junk email which is for no use. But among these spam emails, there is another type of email called phishing email. This phishing email is very dangerous to all internet users especially for multinational companies, finance etc. to everyone who uses even a single account in any of the internet source for various purpose.

Phishing can be defined as an act to steal our valuable information like user id, user password, debit/credit card details for harmful reasons where they are concealed as a genuine organization. Phishing rely on fooling users to share their valuable details regarding usernames, user password, card details etc. phishing can be also defined as a type of cyber-attack that uses electronic communication channels like SMS, emails, phone calls to convey socially manipulated messages to humans which in-turn make them to provide their credentials, credit card number, password etc. for attacker's benefit. Such types of activities persuade a normal website user to enter his/her details to a fraud website that acts like a hidden passage between the user and the attacker. Most of the phishing attacks rely on email and website, that are designed exactly like emails and websites from genuine organization to prompt users into detailing their financial or personal information. The hacker could use this sensitive information of users for his/her benefits.

Many researchers have been working under the phishing problems and proposed a wide variety of solutions to resist phishing attacks. There are two categories regarding the solutions for phishing attacks. In the first category of solutions works by detecting phishing emails or messages to warn the user about the attack before the hacker could steal user's private data. The second category of solutions works by securing the login procedures by adding a secondary login process that will resist the hacker from stealing the credentials.

Word embedding has been a hot topic for language

In: R. Verma, A. Das (eds.): Proceedings of the 1st AntiPhishing Shared Pilot at 4th ACM International Workshop on Security and Privacy Analytics (IWSPA 2018), Tempe, Arizona, USA, 21-03-2018, published at http://ceur-ws.org

identification. Recently, the application of Convolutional Neural Network with Keras embedding is used for e-mail phishing detection [EDB+18]. Following, in this paper, we use Keras Word Embedding and CNN for finding phishing emails from legitimate and phishing ones. Here we aim at developing a classifier which can distinguish phishing emails from legitimate emails. Our model makes use of Keras Word Embedding and Convolutional Neural Network followed by Pooling layer, Fully Connected layer, Non-linear activation function (Sigmoid) and one output neuron for classifying legitimate and phishing mails from the given set of emails.

2 Related Works

In [SAZ+15] Sami S et.al proposed a model for detecting phishing emails that rely on a preprocessing technique which extracts different part of email as feature. And this extracted feature is fed into a j48 classification algorithm to perform classification. In [SZL+15], they considered meaningless tokens and new pages as the feature set. Authors in [SZL+15], selected some features that have better predictability from initial feature set. They provide the O(1) complexity as an evaluation method to each feature set to evaluate its predictive ability. In the paper [KK15], sukhjeel kaui et.al used Genetic algorithm for the detection of phishing webpage and for categorizing pages they preferred a filter function. Lu fang et.al in [FBJ+15] proposes some solution to overcome the time lag in detecting phishing websites. Here they provide a solution to detect phishing websites by analyzing the peculiarity in its WHOIS and URL information. In [VSP18b, VSP18a] deep learning methods were employed to detect malicious URL′s and domains. Binay kumar et.al has used html contents for detecting email phishing in [KKMK15]. But Rachna Dhamija et.al in [TC09] mainly concentrated in this topic to know which phishing activity works during the attack and why. For that they used a large given set of data which contains reported phishing activities. Fergus toolan et.al made a different approach. They used only five features for classification. For classification they used a C5.0 algorithm which have more precision compare to other algorithms. Mayank pandey et.al in [PR12] used different types of classification methods such as Multilayer Perceptron (MLP), Decision Trees (DT), Support Vector Machine (SVM), Group Method of Data Handling (GMDH), Probabilistic Neural Net (PNN), Genetic Programming (GP) and Logistic Regression (LR). Lew may form et.al in [FCT+15] proposed a method which uses hybrid features for detecting phishing emails. It is called Hybrid features because it is a combination of URL based, behavior based and contend based features. Here they acquired an overall accuracy of 97.25 % with an error percentage of 2.75 %. Justin zhan et.al in [ZT11] used a weak estimator method which works by anomaly detection that detects the system which exhibits a deviation in its behavior from the normal system. In [Zen17], they created a machine learning model for detecting phishing emails. Machine learning model was created using a predictive analysis to detect the dissimilarity between both phishing and legitimate emails using a static analysis. Samual marchal et.al in [MFSE14] developed an automatic phishing detecting system that works on real time. This system uses URL that generated from the queries of search engines like yahoo, google etc. for feature extraction. This extracted feature is then used for classification using machine learning. In [FM15], they used host based and lexical features for classifying the URL. They created clusters for the entire dataset which in turn used as a feature for the classification system. This system achieves an accuracy of 93-98 % in detecting phishing emails. Hicham tout in [TH09], done a different approach in which online system should prove their originality for the transfer of data between them.

3 Background

3.1 Keras Word Embedding

Relative meanings and dense representations of words can be provided using word embedding. The sparse representation used by bag of word models are improved using word embedding. In word embedding projection of a word in a continuous vector space is represented by dense vectors. Keras provides an embedding layer which can be used on text data. It requires the input data to be integer encode thus providing a unique integer representation for each word. Initially random weights are assigned to embedding layer which are then modified by learning an embedding for each word in training dataset. It is defined as the first hidden layer of a network. We have to specify three arguments for this layer namely input dimension, output dimension and input length.

3.2 Convolutional Neural Network

Convolutional Neural Networks are several layers of convolutions followed by nonlinear activation function like ReLU. Unlike in traditional neural network where we have fully connected layers, in CNN convolution over input is done to compute the output which results in a local connection. A large number of filters are applied in each layer whose outputs are combined to get the result. Values of filters are learned by CNN during training phase. For NLP tasks the input to CNN will be sentences or documents. A word or a character is

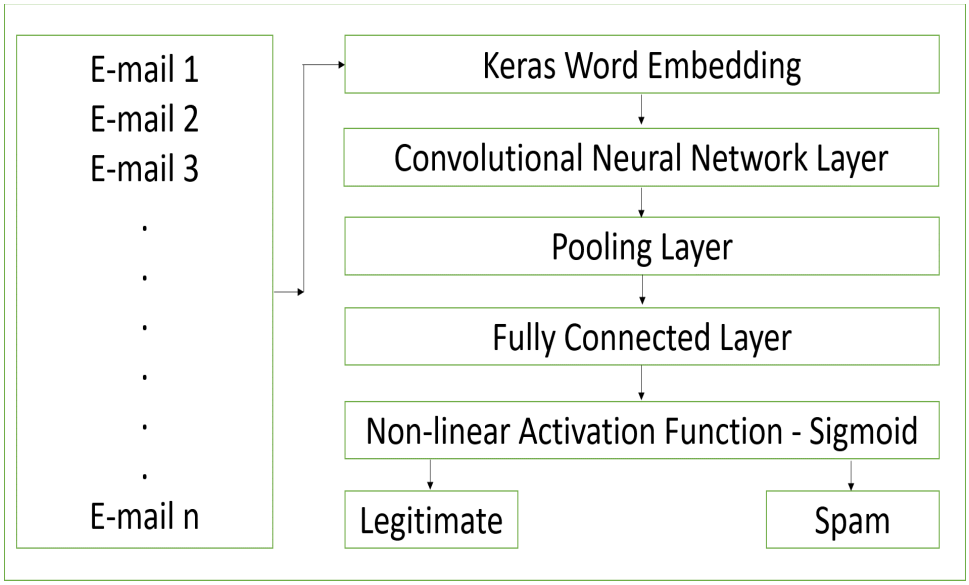

Figure 1: Proposed Architecture

represented as a row of the matrix which provides the vector corresponding to that word known as word embedding. The embedding dimension determines the column space of the matrix. The main difference in CNN between image and NLP is in choosing the size of the filter. In images the filter is slide over a local patch of the input where as in NLP it slides over an entire row since the entire represents a word. In other words column space of filter matrix will be same as column space of input matrix [ZZL15].

4 Experiments

All experiments were run on a GPU enabled TensorFlow [ABC+16] in conjunction with Keras [Cho15] framework. Model was trained using backpropagation methodology. The emails were tokenized and converted to lower case. A dictionary was created which contains a unique id for every word and unknown words were assigned to default key 0. A unique vector is formed for each email and it coordinately works with CNN layer to give a dense vector. We created a total of five models with Keras embedding and CNN layer. Three models for task 1 with CNN epochs varying from 100, 500, 1000. Two models for task 2 with CNN epochs varying from 100, 500.

4.1 Description of Data set

The data set consist of emails having both legitimate and phishing mails [EDMB+18]. Two sets of data sets were given one with header files for Task 1, i.e., having from, to addresses and one without header for Task 2, i.e., only the matter. For training data set, total number of 4,583 mails were given for Task 1 in which 4,082 were legitimate and 501 were phishing. For Task 2, total of 5,700 mails were given in which 5,088 were legitimate while 612 were phishing. For test data set, total of 4,195 emails were given for Task 1 and 4,300 were given for Task 2.

4.2 Proposed Architecture

The Architecture composed of following layers, Keras Embedding, CNN, Classification. Keras embedding is an inbuilt function in Keras framework which generates the vectors for words. A unique vector is formed for each unique words and is then passed to CNN to give a dense vector. The CNN combines the vector formed by embedding layer and gives a much more dense vector which is the passed through pooling layer to reduce the dimensionality and is then given to a fully connected layer. A schematic diagram of the proposed architecture is shown in Figure 1. The model configuration details for both the tasks are given in Table 1. Total parameters for the model is 413105 out of which 413105 are trainable parameters and 0 non-trainable parameters.

5 Results

The model build using the above architecture was used to classify the data set. For sub task 1 in which the emails didn't had header files our model gave an accuracy of 96.8%. For sub task 2 in which header files were given our model gave an accuracy of 94.2 %. The accuracy obtained was measured on a 10 fold cross validation. The results are summarized in Table 2. Our model was tested using test data by IWSPA-AP Shared Task committee and the resulting True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) has been summarized in Table 3.

Table 1: Model Configuration Details

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 1000)	0
embedding_1 (Embedding)	(None, 10000, 100)	4400
conv1d_1 (Conv1D)	(None, 9996, 128)	64128
max_pooling1d_1 (MaxPooling1D)	(None, 1999, 128)	0
conv1d_2 (Conv1D)	(None, 1995, 128)	82048
max_pooling1d_2 (MaxPooling1D)	(None, 399, 128)	0
conv1d_3 (Conv1D)	(None, 395, 128)	82048
max_pooling1d_3 (MaxPooling1D)	(None, 11, 128)	0
flatten_1 (Flatten)	(None, 1408)	0
dropout_1 (Dropout)	(None, 1408)	0
dense_1 (Dense)	(None, 128)	180352
dense_2 (Dense)	(None, 1)	129

Table 2: Cross Validation Results

Method	Task	Accuracy
Word Embedding + CNN	Sub task1 no header	0.968
Word Embedding + CNN	Sub task2 with header	0.942

Table 3: Statistics of Test Result

Method	Task	TP	TN	FP	FN
CNN 100 epochs	No Header	3646	295	180	179
CNN 500 epochs	No Header	3666	288	187	159
CNN 1000 epochs	No Header	3688	287	188	137
CNN 100 epochs	With Header	3237	496	0	462
CNN 500 epochs	With Header	3618	496	0	81

6 Conclusion

Email phishing is a growing threat to digital world. To curb this problem has become a major goal for every digital platform. Here we proposed a model using Keras Word Embedding and CNN to classify legitimate and phishing mails. Combining these two will give a dense vector representation for words which are then used to classify mails given in data set. Our model performed well for both the tasks with header and without header. A highly imbalanced data sets were given for both sub tasks and the task it self was unconstrained, i.e., any data sets can be used during training. But without using any external data sets we were able to get good detection rate for phishing email in both sub tasks. Thus we can conclude that if we add some additional data sources we can considerable increase the detection rate of phishing emails for the proposed methodology.

6.0.1 Acknowledgements

This research was supported in part by Paramount Computer Systems. We are grateful to NVIDIA India, for the GPU hardware support to the research grant. We are grateful to Computational Engineering and Networking (CEN) department for encouraging the research.

References

[ABC+16] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[Cho15] François Chollet. Keras. https://github.com/fchollet/keras, 2015.

[EDB+18] Ayman Elaassal, Avisha Das, Shahryar Baki, Luis De Moraes, and Rakesh Verma. Iwspa-ap: Anti-phising shared task at acm international workshop on security and privacy analytics. In *Proceedings of the 1st IWSPA Anti-Phishing Shared Task*. CEUR, 2018.

[EDMB+18] Ayman Elaassal, Luis De Moraes, Shahryar Baki, Rakesh Verma, and

Avisha Das. Iwspa-ap shared task email dataset, 2018.

[FBJ+15] Lv Fang, Wang Bailing, Huang Junheng, Sun Yushan, and Wei Yuliang. A proactive discovery and filtering solution on phishing websites. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 2348–2355. IEEE, 2015.

[FCT+15] Lew May Form, Kang Leng Chiew, Wei King Tiong, et al. Phishing email detection technique by using hybrid features. In *IT in Asia (CITA), 2015 9th International Conference on*, pages 1–5. IEEE, 2015.

[FM15] Mohammed Nazim Feroz and Susan Mengel. Phishing url detection using url ranking. In *Big Data (BigData Congress), 2015 IEEE International Congress on*, pages 635–638. IEEE, 2015.

[KK15] Sukhjeel Kaui and Amrit Kaur. Detection of phishing webpages using weights computed through genetic algorithm. In *MOOCs, Innovation and Technology in Education (MITE), 2015 IEEE 3rd International Conference on*, pages 331–336. IEEE, 2015.

[KKMK15] Binay Kumar, Pankaj Kumar, Ankit Mundra, and Shikha Kabra. Dc scanner: Detecting phishing attack. In *Image Information Processing (ICIIP), 2015 Third International Conference on*, pages 271–276. IEEE, 2015.

[MFSE14] Samuel Marchal, Jérôme François, Radu State, and Thomas Engel. Phishstorm: Detecting phishing with streaming analytics. *IEEE Transactions on Network and Service Management*, 11(4):458–471, 2014.

[PR12] Mayank Pandey and Vadlamani Ravi. Detecting phishing e-mails using text and data mining. In *Computational Intelligence & Computing Research (ICCIC), 2012 IEEE International Conference on*, pages 1–6. IEEE, 2012.

[SAZ+15] Sami Smadi, Nauman Aslam, Li Zhang, Rafe Alasem, and MA Hossain. Detection of phishing emails using data mining algorithms. In *Software, Knowledge, Information Management and Applications (SKIMA), 2015 9th International Conference on*, pages 1–8. IEEE, 2015.

[SZL+15] Hongzhou Sha, Zhou Zhou, Qingyun Liu, Tingwen Liu, and Chao Zheng. Limited dictionary builder: An approach to select representative tokens for malicious urls detection. In *Communications (ICC), 2015 IEEE International Conference on*, pages 7077–7082. IEEE, 2015.

[TC09] Fergus Toolan and Joe Carthy. Phishing detection using classifier ensembles. In *eCrime Researchers Summit, 2009. eCRIME'09.*, pages 1–9. IEEE, 2009.

[TH09] Hicham Tout and William Hafner. Phishpin: An identity-based anti-phishing approach. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 3, pages 347–352. IEEE, 2009.

[VSP18a] R Vinayakumar, KP Soman, and Prabaharan Poornachandran. Detecting malicious domain names using deep learning approaches at scale. *Journal of Intelligent & Fuzzy Systems*, 34(3):1355–1367, 2018.

[VSP18b] R Vinayakumar, KP Soman, and Prabaharan Poornachandran. Evaluating deep learning approaches to characterize and classify malicious urls. *Journal of Intelligent & Fuzzy Systems*, 34(3):1333–1343, 2018.

[Zen17] Yuanyuan Grace Zeng. Identifying email threats using predictive analysis. In *Cyber Security And Protection Of Digital Services (Cyber Security), 2017 International Conference on*, pages 1–2. IEEE, 2017.

[ZT11] Justin Zhan and Lijo Thomas. Phishing detection using stochastic learning-based weak estimators. In *Computational Intelligence in Cyber Security (CICS), 2011 IEEE Symposium on*, pages 55–59. IEEE, 2011.

[ZZL15] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.