

A Novel Phishing Email Detection Algorithm based on Multinomial Naive Bayes Classifier and Natural Language Processing

Omar Abdelaziz^a, Sahana Deb, Rania Hodhod^b and Lydia Ray^c
TSYS School of Computer Science, Columbus State University, Columbus, GA, U.S.A.

Keywords: Phishing Emails, Naïve Bayes' Classifier, Natural Language Processing.

Abstract: Phishing attacks are a type of social engineering attacks which trick the user into sharing sensitive and personally identifiable information. With the use of machine learning techniques attackers are implementing new methods to scheme more convincing socially engineered messages making it harder for the victims to identify them. With about 3.8 billion email users worldwide and an average person receiving more than 100 emails per day, the importance of efficient and automatic detection of fraudulent emails becomes paramount. This paper proposes a novel application of natural language processing (NLP) and Naïve Bayes' (NB) classifier to identify legitimate and phishing emails. The results show that Bayes' classifier can be used effectively to detect phishing emails with accuracy of 96.03% and 97.21% for balanced and imbalanced datasets, respectively.


1 INTRODUCTION


Emails are one of the most effective methods of communication in the modern world; email or electronic mail finds its use in every field. With increasing popularity and usage, emails are also used for illegal activities such as sending spam or phishing emails. A spam is unsolicited and unwanted junk email sent out in bulk to an indiscriminate recipient list. Typically, spam is sent for commercial purposes. On the other hand, a phishing attack is a type of social engineering attack often used to steal sensitive user data by exploiting certain human psychological weaknesses. Phishing emails exploit problems in everyday human life, such as an overwhelming amount of emails to check, lack of time to scrutinize every detail of an email header or even email language, identical logos of a user's bank or credit card company etc. For example, a phishing email may look like one coming from the user's bank account and may describe some urgent situation to trick the user into clicking a link and giving away her authentication credentials. According to a research conducted in Harvard and Berkley universities, only 23% users go through the content of sites, the links of


which are included in the mails, to decide on their authenticity. (Ma et al. 2009).

Phishing emails are also used to trick users into clicking on a malicious link or attachment in order to deliver malware to a user's system. Many ransomware attacks in recent times were launched by phishing emails (Vinayakumar et al. 2017). There has been a significant increase in the spread of malicious code along with the fraudulent emails, which pointedly targets user account information. (Jason Milletary US-CERT 2005). Once these programs are installed on the victim's computer, they try to retrieve confidential information by infecting various applications and communications (Jason Milletary US-CERT 2005).

With the advancement in technological solutions of cybersecurity, timely patches of applications and operating systems, and awareness of cybersecurity among information technology professionals, penetrating a system or network by exploiting any vulnerability in an application, OS or web application is becoming increasingly harder. A well protected network employs well-configured firewalls with deep packet inspection technology that can block suspicious incoming traffic. The server side attacks,

^a  <https://orcid.org/0000-0001-7694-9114>

^b  <https://orcid.org/0000-0003-3915-8414>

^c  <https://orcid.org/0000-0002-0732-1136>

in which attackers seek vulnerable applications listening to and communicating with outside traffic, require attackers to have increasingly advanced knowledge and technological skills in order to launch a successful attack. In comparison, the client side attacks require attackers to use much cheaper and easier social engineering tricks against general, less cybersecurity aware populations. In client side attacks, attackers send phishing emails with baits for malicious links or attachments to a large population and wait for response. While the downside of such an attack is that attackers have to wait indefinitely for someone to catch their bait, the advantage is that no organizational network, no matter how well protected, is able to block an incoming email. In addition, when an employee from an internal network connects to the attacker's machine by clicking on the malicious link in a phishing email, the outgoing traffic from the victim's computer cannot be blocked either. Therefore, even if there is some wait time in phishing attacks, the chance of success is very high. Due to the cost-effectiveness, phishing attacks are gaining popularity and are used by attackers with a wide variety of motives, ranging from financial gains, blackmail with ransomware, to cyberwarfare by nation-states (Vinayakumar et al. 2017). What's more, phishing attacks are becoming increasingly sophisticated, making it harder for common people to detect the attack.

In our research, we identify the following features of a phishing attack in order to find an effective defense:

- exponentially increasing number of emails flooding to a person's (specifically, a working professional) inbox, making it difficult to manually detect a malicious email;
- increasingly sophisticated language and look of malicious emails, making it harder to detect the deception manually;
- the fact that the content of an incoming email cannot be verified by a firewall;
- the fact that firewalls are unable to block outgoing traffic from a legitimate user's machine.

These trending features of phishing attacks demonstrate that phishing attacks will evolve to be increasingly deceptive and frequent, thereby requiring a more potent defender than just spam filters. With the advancement of data driven artificial intelligence and natural language processing techniques, we envision to build an AI powered defender, which will continuously review emails, expand its knowledge base and will continue to improve its accuracy and false negative rates. This

work uses natural language processing (NLP) techniques including bag-of-words, tokenization, and stemming for lexical and semantic analysis of the emails. The results from applying the NLP techniques are used to train the Bayes' classifier. Bayes' classifier is used to classify the email body. The Naïve Bayes classifier is reasonably robust, fast, and accurate. Most importantly, it is not sensitive to irrelevant features which makes it a suitable choice for email classification.

The next section in the paper describes the related works in this field, Section 3 describes the database used in this work and illustrates the use of NLP and Naïve Bayes classifier to detect phishing emails. Section 4 discusses the results and finally, Section 5 concludes the work.

2 RELATED WORKS

Phishing attacks are growing massively every year (Vinayakumar et al. 2018; Hasan, Hasan & Zahan 2019). Most of the phishing attacks are sent through emails (Vinayakumar et al. 2018; Hiransha et al. 2018). Both machine learning (Miyamoto, Hazeyama & Kadobayashi 2008) and deep learning techniques (Hasan, Hasan & Zahan 2019; Hiransha et al. 2018) have a very satisfactory overall performance as can be seen from the related works reviewed in the rest of this section.

A number of experiments have been done to detect phishing emails using word embedding and neural bag-of-ngrams plus deep learning techniques, including convolutional neural networks (CNN), recurrent neural networks (RNN), long short term memory (LSTM), and multi-layer perceptron (MLP). The experiments were done on two datasets (with and without headers). Results show average accuracy of 97.5% on the no headers dataset and 97% on the headers dataset for all 4 different neural network experiments, (Vinayakumar et al. 2018), while the study done by (Hiransha et al. 2018) using the same dataset to detect phishing emails with using word embedding plus CNN resulted in an accuracy of 96.8% with no-headers and 94.2% with headers.

Machine Learning Approaches have also been used to classify phishing emails using incorporating key structural features and employing different machine learning such as SVM, Neural Networks, Self Organizing Maps and k-Means. SVM showed the best results (Basnet, Mukkamala & Sung 2008).

Natural Language Processing and WordNet were used to detect phishing emails, which do not contain links, but had some other features including

recipients' replies, absence of recipients' names, mentioning money and or currencies, and detecting a state of urgency. The classifier was tested on 1000 emails (600 phishing, 400 legitimate) resulting in accuracy, precision, and recall (sensitivity) of 99% (Aggarwal, Kumar & Sudarsan 2014).

A rule-based approach was used to detect phishing webpages using decision trees and logistic regression learning algorithms. The model has achieved a 95-99% accuracy by using the DT and LR learning algorithms with a false-positive rate of 0.5-1.5% (Basnet, Sung & Liu 2011). Decision trees have proven to be inexpensive and fast at classification tasks. However, they could be difficult to interpret. Also, small changes in the input data may result in large changes in the decision. Decision trees showed the lowest false-positive rate in comparison to Random Forest, K Nearest Neighbor, Logistic model tree, and J48 (Alhawi, Baldwin & Dehghantanha 2018).

As can be seen from the literature review done and related works, a number of deep learning and machine learning techniques were used to detect phishing emails including neural networks and natural language processing, rule-based approach, decision trees, and support vector machine (SVM). To the extent of our knowledge, there is no research done on the use of Naïve Bayes classifier to detect and classify phishing emails. This paper proposes a new approach to detect phishing emails using bag-of- words and Naïve Bayes' classifier.

3 DETECTION OF PHISHING EMAILS

This work proposes a novel way of applying Naïve Bayes' classifier to detect phishing emails. Figure 1 shows the flow diagram of the classification process.

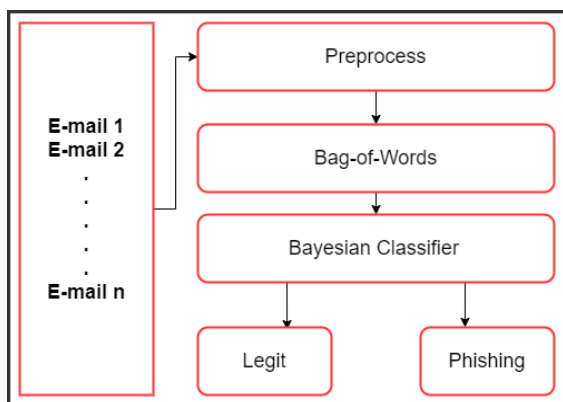


Figure 1: Flow of the classification process.

The following subsections provide an illustration of each of the steps in the flow diagram.

3.1 Dataset Description

In this work, we have used the IWSPA dataset (Verma, Zeng, & Faridi 2019), which consists of two sets of emails (with headers and without headers) under two categories (legit and phishing). There are around 5,719 (5091 legit, 628 phishing) emails without headers. Whereas the other dataset has 4,082 legit emails and 503 phishing emails with headers.

3.1.1 Data Preprocessing

Lexical and semantics analysis of the IWSPA dataset has been used to extract the most frequently used words in the phishing and legit emails. The dataset was preprocessed first which includes removal of special characters, single characters and multiple spaces, lemmatization and converted into lowercase characters. The bag-of-words model is then implemented and the results from the model were passed to the multinomial Naïve Bayes' classifier.

3.1.2 NLP and Bag-of-Words

We secondly implement the bag-of-words model for feature generation. We obtain a vector representation of the term frequencies of the fifty most frequent terms in our dataset. The dataset that has been used is extracted from the series of fraudulent emails. We begin with preprocessing the text, which includes removal of non-alphabetic characters, conversion of the text into lower case, removing of stop words, and stemming the data. We filter out the fifty most frequent words from the data which are fed to the bag of words model. The output is a matrix representing the term frequencies as they appear throughout the data.

The bag-of-words provided a number of significant words in phishing email headers like Table, Message, Bank, Security, Delayed, Account, HTTP, Abuse, Please, SPAM, eBay, PayPal; phishing email bodies like Account, eBay, PayPal, link, Please, Privacy, Password, banking, money, buy; legitimate email headers like Unknown, Customer, Pascal, Encrypt, Trump, United, Twitter, Deadline, Spam, Amazon; and legitimate email bodies like Trump, Donald, Walker, Daily, Rebecca, Republican, Democrat, political, Obama, United, States, Candidate.

By eliminating the common words between phishing and legitimate emails, we obtain a unique set of words that are in the phishing emails only. This includes: PayPal, eBay, Delay, Banking, Invoke, Exchange, Chase, SPAM, Unpaid, eBay, PayPal,

Bank, Security, Please, update, Privacy, Copyright, Click, Help

3.1.3 Naïve Bayes Classifier

Naïve Bayes classifier is a probabilistic machine learning algorithm that can be used in a wide variety of classification tasks (Feng et al. 2016). It has been proven effective in detecting phishing websites (James, Sandhya & Thomas 2013), spam filtering (Feng et al. 2016) and detecting phishing emails (Yasin, Abuhasan 2016). In Yasin, Abuhasan (2016) work, the Naïve Bayes’ classifier shows precision and recall of 94%.

A multinomial Naïve Bayes classifier is applied to the IWSPA dataset into a ratio of 75:25 for training and testing purposes respectively.

As described in Section 3.3.1, the dataset is imbalanced so in addition to applying the classifier on the imbalanced dataset, we have also created another balanced dataset which includes 503 emails for each of the legit and phishing classes. The following calculations have been applied:

$$\text{Accuracy} = (\text{True Negative} + \text{True Positive}) / \text{Total Dataset} \quad (1)$$

$$\text{Misclassification Rate} = (\text{False Negative} + \text{False Positive}) / \text{Total Dataset} \quad (2)$$

$$\text{Precision} = \text{True Positive} / (\text{True Positive} + \text{False Positive}) \quad (3)$$

$$\text{F-Score} = 2 * (\text{Sensitivity} * \text{Precision}) / (\text{Sensitivity} + \text{Precision}) \quad (4)$$

$$\text{Sensitivity} = \text{True Positive} / (\text{True Positive} + \text{False Negative}) \quad (5)$$

$$\text{Specificity} = \text{True Negative} / (\text{True Negative} + \text{False Positive}) \quad (6)$$

The classification results from both datasets are listed in Table 1.

Table 1: Results of both Balanced and Imbalanced Datasets.

Variable / Dataset	Balanced	Imbalanced
Accuracy	96.03%	97.21%
Miss Classification	3.96%	2.78%
Precision	95.27%	83.00%
F-Score	96.03%	88.81%
Sensitivity	96.80%	95.48%
Specificity	95.27%	97.43%
Ture Positive	121	127
True Negative	121	988
False Positive	6	26
False Negative	4	6

The “Positive” values indicate phishing emails and “Negative” values represent legitimate emails while “True” acts of the right classified emails and “False” represent the incorrectly classified emails.

4 DISCUSSION OF RESULTS

NLP techniques have been applied to the IWSPA dataset to extract the significant worlds for both phishing and legit emails. The results have been fed to Naïve Bayes classifier which has been trained twice; one time on the imbalanced dataset and another time on the balanced one. The classifier recorded accuracies, see (1), of 96.03% and 97.21% for both imbalanced and balanced datasets respectively. It is also important to assess the performance of the classifier with respect to other factors including misclassification rate, F-Score, sensitivity, and specificity.

Misclassification is the total number of incorrectly classified emails by the classifier, meaning the summation of false results, see (2). Misclassification rate was less than 4% for both datasets. The classifier has recorded a high F-score, see (4), which shows the effectiveness of the classifier with high precision and recall values. The classifier sensitivity ranges between 96.8% for balanced dataset and 95.48% for the imbalanced dataset. This can be attributed to the large number of legitimate emails in the imbalanced dataset compared to the phishing emails, see (5). Specificity, see (6), is the percentage of the correctly classified emails as legitimate among all legitimate emails in the dataset. Specificity ranged between 95.27% for the balanced dataset and 97.43% for the imbalanced dataset. This can also be attributed to the large number of legitimate emails in the imbalanced dataset compared to the phishing emails,

A comparison had been made between the results achieved in this work and existing works that used the IWSPA dataset. Hiransha et al. (2018) applied Convolution Neural Network technique and achieved 94.2% accuracy rate, while Vinayakumar et al. (2018) applied word embedding plus long short term memory (LSTM) which gave an accuracy of 86% while the Bayes’ classifier used in this work achieved accuracy of 97.4%. On the other hand, the Bayes’ classifier produced a lower F-Score than LSTM, 88.8% and 92.2%, respectively. It is also worth mentioning that Bayes’ classifier achieved an F-Score of 96.03% for a balanced dataset.

Other works used NLP and WordNet recorded higher accuracy of 99% when applied to other datasets (Aggarwal, Kumar & Sudarsan 2014). In

Yasin, Abuhasan (2016) work, the Naive Bayes' classifier shows precision and recall of 94% when applied to a dataset from the Nazario phishing corpus which consists of 5940 legit and 4598 phishing emails.

5 CONCLUSION AND FUTURE WORK

Phishing emails are tricking recipients to either give their credentials (i.e. bank account number, passwords) or force them unconsciously to click on a malicious link which will do nothing but harm to the receiver's computer.

This work proposed a new approach to applying NLP and Naïve Bayes' classifier to detect phishing emails. Based on the achieved results, the classifier offers a higher accuracy than other works that used the same dataset in the literature that is 97.21%. Future work includes the application of the Apriori algorithm to find associations between the significant words in phishing emails. Furthermore, combining all algorithms and testing the new model on the dataset without headers is on our future work plans.

REFERENCES

- Aggarwal, S., Kumar, V., & Sudarsan, S. D. (2014, September). Identification and detection of phishing emails using natural language processing techniques. In Proceedings of the 7th International Conference on Security of Information and Networks (pp. 217-222).
- Ahmed, I., Guan, D., & Chung, T. C. (2014). Sms classification based on naive bayes classifier and apriori algorithm frequent itemset. *International Journal of machine Learning and computing*, 4(2), 183.
- Alhawi, O. M., Baldwin, J., & Dehghantaha, A. (2018). Leveraging machine learning techniques for windows ransomware network traffic detection. In *Cyber Threat Intelligence* (pp. 93-106). Springer, Cham.
- Basnet, R., Mukkamala, S., & Sung, A. H. (2008). Detection of phishing attacks: A machine learning approach. In *Soft computing applications in industry* (pp. 373-383). Springer, Berlin, Heidelberg
- Basnet, R. B., Sung, A. H., & Liu, Q. (2011). Rule-based phishing attack detection. In Proceedings of the International Conference on Security and Management (SAM) (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
- Feng, W., Sun, J., Zhang, L., Cao, C., & Yang, Q. (2016, December). A support vector machine based naive Bayes algorithm for spam filtering. In 2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC) (pp. 1-8). IEEE.
- Hasan, K. Z., Hasan, M. Z., & Zahan, N. (2019, July). Automated Prediction of Phishing Websites Using Deep Convolutional Neural Network. In International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE.
- Hiransha, M., Unnithan, N. A., Vinayakumar, R., Soman, K., & Verma, A. D. R. (2018, March). Deep learning based phishing e-mail detection. In Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal. (IWSPA). Tempe, AZ, USA.
- James, J., Sandhya, L., & Thomas, C. (2013, December). Detection of phishing URLs using machine learning techniques. In 2013 international conference on control communication and computing (ICCC) (pp. 304-309). IEEE.
- Jason Military. (2005, May). Technical trends in Phishing attacks
- Ma, L., Ofoghi, B., Watters, P., & Brown, S. (2009, July). Detecting phishing emails using hybrid features. In 2009 Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing (pp. 493-497). IEEE.
- Miyamoto, D., Hazeyama, H., & Kadobayashi, Y. (2008, November). An evaluation of machine learning-based methods for detection of phishing sites. In International Conference on Neural Information Processing (pp. 539-546). Springer, Berlin, Heidelberg.
- Verma, R. M., Zeng, V., & Faridi, H. (2019, November). Data Quality for Security Challenges: Case Studies of Phishing, Malware and Intrusion Detection Datasets. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (pp. 2605-2607).
- Vinayakumar, R., Barathi Ganesh, H. B., Anand Kumar, M., & Soman, K. P. (2018). Deep Anti-PhishNet: Applying Deep Neural Networks for Phishing Email Detection. *CEN-AISecurity@IWSPA*, 40-50.
- Vinayakumar, R., Soman, K. P., Velan, K. S., & Ganorkar, S. (2017, September). Evaluating shallow and deep networks for ransomware detection and classification. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 259-265). IEEE.
- Yasin, A., & Abuhasan, A. (2016). An intelligent classification model for phishing email detection. arXiv preprint arXiv:1608.02196.