Spam Email Detection: Comparison Between Naïve Bayes and Neural Network

Zhuolin Li, Hao Zhang
Department of Computer Science
Kennesaw State University
Marietta, USA
{zli29,
hzhang13}@students.kennesaw.edu

Mohammad Masum
Analytics and Data Science Institute
Kennesaw State University
Marietta, USA
mmasum@students.kennesaw.edu

Hossain Shahriar, Hisham Haddad
Department of Information Technology
Department of Computer Science
Kennesaw State University
Marietta, USA
hhaddad@students.kennesaw.edu

Abstract- *Classification is an important technique to deal with cybersecurity threats. In this paper, we detect spam emails from publicly available dataset using Naive Bayes and Neural Network (NN). The results from experiments show that for data sets with more balanced for classification, the accuracy of Naive Bayes is better than NN.*

Keywords: Spam email, Naive Bayes, Neural Network.

I. INTRODUCTION

Email is an efficient mode of online communication because it saves an expense and it helps in the reduction of time which it takes for effective communication to occur and as a result this makes it a favorable means of communication both on personal and business perspectives. Spam emails on the other hand is the practice of sending and flooding a server with unwanted and huge amount of data which is targeted to specific email accounts and this emails usually includes malwares which are in the form of scripts or various other files that can be executed in a certain way with the sole purpose of harming a user's system [1, 5].

The cost implication of spam emails can be very huge to every organization and this is the reason why it is very important to be able to identify and classify which emails are spam [17]. Supervised machine learning methods for classifying spam emails were established long ago. Most of these methods nowadays either use header-based or content-based features.

Security threats are evolving and getting more hidden and complicated. Detecting malicious security threats and attacks have become a huge burden to our cyberspace. We should apply proactive prevention and early detections of security vulnerabilities and threats rather than patching security holes afterwards. To analyze the huge amount of data to find out suspicious behaviors, threat patterns, and vulnerabilities and to predict and prevent future cybersecurity threats are a challenge. Machine Learning (ML) is a powerful instrument to take up such challenge.

In this paper, we apply Naïve Bayes and Neural Network (NN) classifier to detect spam emails with imbalanced dataset. Balanced accuracy is calculated as the average of the proportion corrects of each class individually. We apply the balance accuracy to analysis the performance of the three supervised algorithms on imbalanced dataset. The initial results show that Naïve Bayes perform better than NN.

The rest of this article is organized as follows. Section II briefly introduces the related literature work. Section III introduces the dataset we used in our experiment. Section IV discusses the classifier techniques and the research results. Finally, Section V concludes the paper.

II. RELATED WORK

Chan et al. [19] applied Naive Bayes classifier to combat spam email attack, where each feature in the Naive Bayes classifier, additional weight based on the number of ham and spam containing the feature is added. Support vector machines has been applied for the classification of spam emails [20]. Case Base Spam Filtering has been proposed that include pre-processing, feature extraction, and selection, grouping of email data [21]. Heuristic-based Filtering Technique uses already created rules or heuristics to assess a huge number of patterns which are usually regular expressions against a chosen message [22]. Other efforts include comparison between algorithms for classification problems [5, 10, 11, 12, 13, 14, 15, 16] along with various measures of performance namely precision, recall, f-measure and accuracy. In contrast, this work compared Naïve Bayes with NN for spam email detection with balanced accuracy measure.

III. DATASETS

Our dataset comes from Kaggle. The goal of the dataset is to judge if the emails are spams or hams [3]. There are 5,728 emails (4,360 hams and 1,368 spams) in this data set. We split the dataset into training and test sets. The training set includes 4,296 examples (about 75% of the whole data set), while the rest of data is named as test set: 1432 examples (about 25% of the whole data set).

> Subject: 4 color printing special request additional information now ! click here click here for a printable version of our order form (pdf format) phone : (626) 338 - 8090 fax : (626) 338 - ...

Figure 1. Sample for spam email

For the email dataset, we list 2 samples in Figures 1 (spam) and 2 (ham). As the dataset uses text, we need do some data

978-1-7281-6861-6/20/$31.00 ©2020 IEEE

pre-processing. First, we split the entire text into sentences and then, the sentences into words. We then change the words in lowercase and eliminate all the punctuation on the text.

> Subject: re : london contact number hi anita , how are you ? i arrived yesterday late morning from the london gatwick airport . due to rush hour traffic , etc . it took a while to get into the city...

Figure 2. Sample for ham email

Then, we use a function named *nltk.stemming* from python text mining library to normalize the text. For example, we classify word "take", "takes", "took", "token" into one word "take". We use Porter Stemmer from python text mining library to let pluralized words into its corresponding single version. We divide both the datasets into training set and test set respectively. The training set of the two datasets accounts for 75% of each total sample, and the test set accounts for 25% of each total sample. And then we also test 20% (Testing) / 80% (Training) and 30% (Testing) / 70% (Training).

IV. METHODOLOGY AND RESULTS

A. Naive Bayes

Naïve Bayes classification usually adopts the strategy of content-based filtering technique [4, 18]. This method analyses words, the occurrence, distributions of words and phrases in the content of emails and then use generated rules to filter all the incoming spam emails. It can be further illustrated as an approach which is based on a statistical machine learning process which has the properties of an independence which is strong and equally can handle a large number of datasets. In the concept of Naïve Bayes, the distribution of a probability is usually assessed from the rate of distribution of the dataset. The calculation of the probability of a spam email using the Naive Bayes methodology can be described as below:

P (spam word) = P(Spam). P (word spam)/P(spam). P (word spam) +P(non-spam). P(word|non-spam).

B. Neural Network (NN)

Neural Network (NN) (Figure 3) consist of three layers: input layer, hidden layers (often more than two) and output layer [2, 7]. Each layer is made up of nodes. At the input layer, we convert the original data into numbers for input, multiply them by their corresponding weights, add them up to the value of the node, and apply activation function after each node is calculated. The main purpose of activation function is to convert the input signal of a node in the neural network into the output signal. This output signal is used as an input to the next layer. A hidden layer is added between the input and output layers to amplify the function of the neural network and improve its accuracy. However, unlimited adding hidden layer may lead to the increase of computation time and the decrease of accuracy.

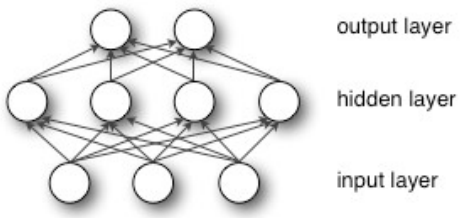

Figure 3 Artificial Neural Network

For our experiment, we have 3 hidden layers with 20 nodes respectively. And the output layer has 2 output because our classification problem is binary. The activation function determines the output, based on its input. We usually use Relu function (Eq. 1) in our Hidden layer, and the softmax function (Eq. 2) to our output layer.

$$f(x) = \max(0, x) \tag{1}$$

$$f(x) = \frac{e^{x_i}}{\sum_k e^{x_k}} \tag{2}$$

V. EVALUATION

We evaluate the data set with the two algorithms. For the data set of spam email detection, true positive means that positive examples are correctly assigned to the positive class. In this data set, it means this email is ham. True negative refers to the negative examples correctly predicted to the the negative class. It means this email is spam.

False positive is related to the algorithm is wrong to consider negative examples as positive examples. It means the sample email is spam. However, the algorithm mistakenly placed the spam in the category of ham. False negative is defined as positive examples incorrectly classified to negative class. It means that a sample email is ham, but the algorithm incorrectly classified as spam. We used balanced accuracy [6] to determine whether an algorithm is a good algorithm.

$$\text{Balanced accuracy} = ((TP/(TP + FP)) + (TN/(TN + FN)))/2$$

In the above equation, TP is true positive, FP is false positive. TN is True Negative (TN), FN is False Negative (FN). The higher the balance accuracy is, the more the classification is put into the right place. The balanced accuracy analysis is shown in Table 1. Here, we find that Naïve Bayes outperforms Neural Network.

Table 1: Balanced Accuracy-based comparison between Naïve Bayes and Neural Network.

Dataset split	Naive Bayes	Neural Network
(75%/25%)	0.90600293	0.88324189
(70%/30%)	0.90772206	0.82304718
(80%/20%)	0.91497069	0.84149622

978-1-7281-6861-6/20/$31.00 ©2020 IEEE

VI. CONCLUSION

In this paper, we used a spam email dataset to classify emails by using Naive Bayes and Neural Network. The experimental results show that the Naive Bayes algorithm performed better than logistic regression and neural networks in the dataset with highly imbalanced distribution.

REFERENCE:

[1] L.F. Cranor, B.A. Lamacchia, "Spam!" *Communications of the ACM*, vol. 41, 1998.

[2] Chunhui Bao, Yifei Pu, and Yi Zhang, "Fractional-Order Deep Backpropagation Neural Network," *Computational Intelligence and Neuroscience*, Vol. 2018, Article ID 7361628, 2018.

[3] Karthickveerakumar. (2017). Spam Filter: Identifying spam using, https://www.kaggle.com/karthickveerakumar/spam-filter

[4] Xu, Shuo & Li, Yan & Zheng, Wang. (2017). Bayesian Multinomial Naïve Bayes Classifier to Text Classification. 347-352. 10.1007/978-981-10-5041-1_57.

[5] Wardani, Dewi, et al. "Using Metadata in Detection Spam Email with Pornography Content," *Proc. of International Conference on Electrical Engineering and Computer Science (ICECOS)*, October 2018, pp. 213–218.

[6] Stiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas. "Handling imbalanced datasets: A review". *GESTS International Transactions on Computer Science and Engineering*, Vol.30, 2006

[7] Beginner's Guide to Neural Networks and Deep Learning, Pathmind, skymind.ai/wiki/neural-network.

[8] Paul, A. J., Collins, P. J., & Temple, M. A. (2019). Enhancing Microwave System Health Assessment Using Artificial Neural Networks. *IEEE Antennas and Wireless Propagation Letters,* 18(11), 2230–2234.

[9] Wolpert DH, "The lack of a priori distinctions between learning algorithms," *Neural Computation*, 1996, 8(7), pp. 1341–90.

[10] Wu, Xindong, and Vipin Kumar. The Top Ten Algorithms in Data Mining. CRC Press, 2009.

[11] Marianingsih, Susi, and Fitri Utaminingrum. "Comparison of Support Vector Machine Classifier and Naïve Bayes Classifier on Road Surface Type Classification." *2018 International Conference on Sustainable Information Engineering and Technology (SIET),* Nov. 2018, pp. 48–53.

[12] Princy, G., P. V., "Machine learning approach for filtering spam emails," Proceedings of the 8th International Conference on Security of Information and Networks, pages 271-274. Sochi, Russia, September 2015.

[13] Rahman, A, "Filtering Spam Using Naïve Bayes," 2019, https://towardsdatascience.com/spam-filtering-using-naive-bayes-98a341224038

[14] Nurul Fitriah Rusland, Norfaradilla Wahid, Shahreen Kasim,Hanayanti Hafit, "Analysis for Naive Bayes Algorithm in Email Spam Filter which is among Multi numbers of Datasets," 2017, IOP Conference Series: Materials Science and Engineering, 226, 012091.

[15] Shams, R & Mercer R. (2013). Classifying spam emails using text and readability features. 2013 IEEE 13th International Conference on Data Mining, Dallas, TX, 2013, pp. 657-666.

[16] S. R. Gomes et al., "A comparative approach to email classification using Naive Bayes classifier and hidden Markov model," *Proc. of 4th International Conference on Advances in Electrical Engineering (ICAEE)*, Dhaka, 2017, pp. 482-487

[17] Emilio Ferrara, The History of Digital Spam, *Communications of the ACM*, August 2019, Vol. 62 No. 8, pp. 82-91.

[18] K. Larsen, Generalized Naive Bayes Classifiers, *ACM SIGKDD Explorations Newsletter*, Vol.7, No.1, pp. 76-81.

[19] Juayan, P. & Chan, P. (2019). Revised Naïve Bayes classification for bating the attack in filtering, *2013 International Conference on Machine Learning and Cybernetics*, Tianjin, 2013, pp. 610-614.

[20] E. Dada, et al. "Machine Learning to email spam filtering where review, approaches and open research problems," 2019. https://www.sciencedirect.com/science/article/pii/S2405844018353404

[21] V. Christina, S. Karpagavalli, G. Suganya, Email spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng., 02 (09) (2010), pp. 3126-3129.

[22] J.R. Mendez, F. Díaz, E.L. Iglesias, J.M. Corchado, "A comparative performance study of feature selection methods for the anti-spam filtering domain, Advances in Data Mining," *Applications in Medicine, Web Mining, Marketing, Image and Signal Mining*, Springer, Heidelberg, 2006, pp. 106-120.

978-1-7281-6861-6/20/$31.00 ©2020 IEEE