ISSN: 2278-4632
Vol-10 Issue-6 No. 1 June 2020

Detection of Phishing Attacks using Natural Language Processing and Logistic Regression Model

Rajavardhan Reddy Marikanti[1], Katkoori Shiva Prasad[2], Hannoop Kumar Suddala[3], K. Bala Thripura Sundari[4]

[1,2,3]UG Student, [4]Assistant Professor, Department of CSE
[1,2,3,4]Kommuri Pratap Reddy Institute of Technology, Ghatkesar, Hyderabad, India.

ABSTRACT

The increasing volume of unsolicited bulk e-mail (also known as spam) has generated a need for reliable anti-spam filters. Machine learning techniques now days used to automatically filter the spam e-mail in a phenomenally successful rate. In this paper we presented the most popular machine learning method named as logistic regression and its applicability to the problem of spam Email classification. Further, evaluation of proposed machine learning model is compared to existing K-nearest neighbour (KNN) classifier.

Keywords: Spam detection, E-mail, machine learning, logistic regression.

1. INTRODUCTION

Recently unsolicited commercial / bulk e-mail also known as spam, become a big trouble over the internet. Spam is waste of time, storage space and communication bandwidth. The problem of spam e-mail has been increasing for years. In recent statistics, 40% of all emails are spam which about 15.4 billion email per day and that cost internet users about $355 million per year. Automatic e-mail filtering seems to be the most effective method for countering spam now and a tight competition between spammers and spam-filtering methods is going on. Only several years ago most of the spam could be reliably dealt with by blocking e-mails coming from certain addresses or filtering out messages with certain subject lines. Spammers began to use several tricky methods to overcome the filtering methods like using random sender addresses and/or append random characters to the beginning or the end of the message subject line [11]. Knowledge engineering and machine learning are the two general approaches used in e-mail filtering. In knowledge engineering approach a set of rules must be specified according to which emails are categorized as spam or ham. A set of such rules should be created either by the user of the filter, or by some other authority (e.g. the software company that provides a rule-based spam-filtering tool). By applying this method, no promising results shows because the rules must be constantly updated and maintained, which is a waste of time and it is not convenient for most users. Machine learning approach is more efficient than knowledge engineering approach; it does not require specifying any rules [4]. Instead, a set of training samples, these samples is a set of pre classified e-mail messages. A specific algorithm is then used to learn the classification rules from these e-mail messages. Machine learning approach has been widely studied and there are lots of algorithms can be used in e-mail filtering.

ISSN: 2278-4632
Vol-10 Issue-6 No. 1 June 2020

2. REALTED WORK

There is some research work that apply machine learning methods in e-mail classification, In [2], authors demonstrated that the naïve Bayes e-mail content classification could be adapted for layer-3 processing, without the need for reassembly. Suggestions on pre-detecting e-mail packets on spam control middleboxes to support timely spam detection at receiving e-mail servers were presented. In [1], hardware architecture of na¨ıve Bayes inference engine for spam control using two class e-mail classification is presented. That can classify more 117 million features per second given a stream of probabilities as inputs. This work can be extended to investigate proactive spam handling schemes on receiving e-mail servers and spam throttling on network gateways.

Author in [3] proposed a system that used the SVM for classification purpose, such system extract email sender behavior data based on global sending distribution, analyse them and assign a value of trust to each IP address sending email message, the Experimental results show that the SVM classifier is effective, accurate and much faster than the Random Forests (RF) Classifier.

Author in [11] developed personalized email prioritization (PEP) method that specially focus on analysis of personal social networks to capture user groups and to obtain rich features that represent the social roles from the viewpoint of particular user, as well as they developed a supervised classification framework for modelling personal priorities over email messages, and for predicting importance levels for new messages.

An immune-inspired model, named innate and adaptive artificial immune system (IA-AIS) is presented in [4] and applied to the problem of identification of unsolicited bulk e-mail messages (SPAM). It integrates entities analogous to macrophages, B and T lymphocytes, modelling both the innate and the adaptive immune systems. An implementation of the algorithm could identify more than 99% of legitimate or SPAM messages parameter configurations. It was compared to an optimized version of the naive Bayes classifier, which have been attained extremely high correct classification rates. It has been concluded that IA-AIS has a greater ability to identify SPAM messages, although the identification of legitimate messages is not as high as that of the implemented naive Bayes classifier.

What is Natural Language Processing?

Natural Language Processing, usually shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable. Most NLP techniques rely on machine learning to derive meaning from human languages. In fact, a typical interaction between humans and machines using NLP could go as follows:

- A human talk to the machine.
- The machine captures the audio.
- Audio to text conversion takes place.

ISSN: 2278-4632
Vol-10 Issue-6 No. 1 June 2020

- Processing of the text's data.
- Data to audio conversion takes place.
- The machine responds to the human by playing the audio file.

NLP entails applying algorithms to identify and extract the natural language rules such that the unstructured language data is converted into a form that computers can understand. When the text has been provided, the computer will utilize algorithms to extract meaning associated with every sentence and collect the essential data from them. Sometimes, the computer may fail to understand the meaning of a sentence well, leading to obscure results. For example, a humorous incident occurred in the 1950s during the translation of some words between the English and the Russian languages.

Here is the biblical sentence that required translation:

"The spirit is willing, but the flesh is weak."

Here is the result when the sentence was translated to Russian and back to English:

"The vodka is good, but the meat is rotten."

Here are some syntax techniques that can be used for NLP tasks:

Lemmatization: It entails reducing the various inflected forms of a word into a single form for easy analysis.

Morphological segmentation: It involves dividing words into individual units called morphemes.

Word segmentation: It involves dividing a large piece of continuous text into distinct units.

Part-of-speech tagging: It involves identifying the part of speech for every word.

Parsing: It involves undertaking grammatical analysis for the provided sentence.

Sentence breaking: It involves placing sentence boundaries on a large piece of text.

Stemming: It involves cutting the inflected words to their root form.

3. MACHINE LEARNING IN E-MAIL CLASSIFICATION

Machine learning field is a subfield from the broad field of artificial intelligence, this aims to make machines able to learn like human. Learning here means understood, observe, and represent information about some statistical phenomenon. In unsupervised learning one tries to uncover hidden regularities (clusters) or to detect anomalies in the data like spam messages or network intrusion. In e-mail filtering task some features could be the bag of words or the subject line analysis. Thus, the input to e-mail classification task can be viewed as a two-dimensional matrix, whose axes are the messages and the features. E-mail classification tasks are often divided into several sub-tasks. First, Data collection and representation are mostly problem specific (i.e. e-mail messages), second, e-mail feature

selection and feature reduction attempt to reduce the dimensionality (i.e. the number of features) for the remaining steps of the task. Finally, the e-mail classification phase of the process finds the actual mapping between training set and testing set.

3.1. K-Nearest Neighbour

K-Nearest neighbour is a lazy learner technique. This algorithm depends on learning by analogy. It is a supervised classification method. This classifier is used extensively for classification purpose. This classifier waits till the last minute prior to build some model on a specified tuple as compared to earlier classifiers. The training tuples are characterized in N-dimensional space in this classifier. This classification model looks for the k training tuples nearest to the indefinite sample in case of an indefinite tuple. Then, this classifier puts the sample in the closest class.

Disadvantages

Results with less accuracy as low as 50% due to following:

- **Does not work well with large dataset:** In large datasets, the cost of calculating the distance between the new point and each existing point is huge which degrades the performance of the algorithm.
- **Does not work well with high dimensions:** The KNN algorithm does not work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension.
- **Need feature scaling:** We need to do feature scaling (standardization and normalization) before applying KNN algorithm to any dataset. If we do not do so, KNN may generate wrong predictions.
- **Sensitive to noisy data, missing values, and outliers:** KNN is sensitive to noise in the dataset. We need to manually impute missing values and remove outliers.

3.2. Logistic Regression

Logistic regression is named for the function used at the core of the method, the logistic function. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

$$\frac{1}{(1 + e^{-value})}$$

Where e is the base of the natural logarithms (Euler's number or the exp () function in your spreadsheet) and value is the actual numerical value that you want to transform.

Advantages

- It performs well when the dataset is linearly separable.
- This is less prone to over-fitting, but it can overfit in high dimensional datasets. You should consider Regularization (L1 and L2) techniques to avoid over-fitting in these scenarios.

- It is not only giving a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative).
- This is easier to implement, interpret and very efficient to train.

3.3. Algorithm

Step 1: Email pre-processing

The content of email is received through our software, the information is extracted then as mentioned above, then the information (Feature) extracted is saved into a corresponding database. Every message was converted to a feature vector with 21700 attributes (this is approximately the number of different words in all the messages of the corpus). An attribute n was set to 1 if the corresponding word was present in a message and to 0 otherwise. This feature extraction scheme was used for all the algorithms.

Step 2: Description of the feature extracted

Feature extraction module extract the spam text and the ham text, then produce feature dictionary and feature vectors as input of the selected algorithm, the function of feature extraction is to train and test the classifier [9]. For the train part, this module account frequency of words in the email text, we take words which the time of appearance is more than three times as the feature word of this class. And denote every email in training as a feature vector.

Step 3: Spam classification

Through the steps above, we take standard classification email documents as training document, pre-treatment of email, extract useful information, save into text documents according to fix format, split the whole document to words, extract the feature vector of spam document and translate into the form of vector of fix format. We look for the optimal classification using the selected algorithm which is constructed using the feature vector of spam documents.

Step 4: Performance evaluation

To test the performance of above mentioned six methods, we used the most popular evaluation methods used by the spam filtering researchers.

4. EXPERIMENTAL ANALYSIS

To test the performance of above mentioned six methods, some corpora of spam and legitimate emails had to be compiled; there are several collections of email publicly available to be used by researchers. SpamAssassin (http://spamassassin.apache.org) will be used in this experiment, which contains 6000 emails with the spam rate 37.04%. Thus, we have divided the corpora into training and testing sets keeping, in each such set, the same proportions of ham (legitimate) and spam messages as in the original example set. Each training set produced contained 62.96% of the original set, while each test set

contain 37.04%. In addition to the body message of an email, an email has another part called the header. The job of the header is to store information about the message, and it contains many fields like the field (From) and (Subject), we decided to divide the email into 3 different parts. The first part is the (Subject) that can be considered as the most important part in the email, it noticed that most of the new incoming emails have descriptive Subjects that can be used to clearly identify whether that email is Spam or Ham. The second part is (From) which is the person that taking the responsibility of the message, this field we store it in a database and use it after the decision of the classifier has been taken, that is the way to compare the field (From) stored in the database to the field (From) in the new incoming email, if they are the same so the decision of the new incoming email is Spam. The (Body) is the third part which is the main part of the message. Furthermore, we applied two procedures in the pre-processing stage. Stopping is employed to remove common word. Case-change is employed to change the (Body) into small letters. The experiment is performed with the most frequent words in spam email; we select 100 of them as features.

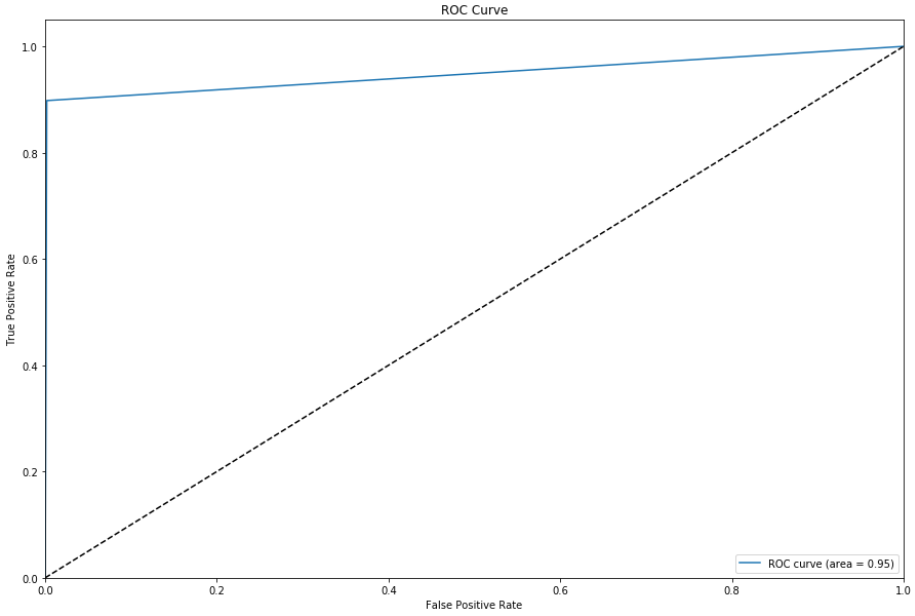

Fig. 1: ROC curve.

Table 1. Quality evaluation.

	Accuracy (in %)	Precision (in %)	Recall (in %)
KNN classifier	50	68	70
Logistic regression	96	87	91

The performance of the KNN classifier appeared to be nearly independent of the value of k. In general, it was poor, and it has the worst precision percentage. The performance of the logistic regression is the most simple and fastest algorithm.

5. CONCLUSIONS

In this paper we implemented the most popular machine learning method and its applicability to the problem of spam e-mail classification. Description of the algorithms are presented, and the comparison of their performance on the SpamAssassin spam corpus is presented, the experiment showing a very promising results specially in the algorithms that is not popular in the commercial e-mail filtering packages, spam recall percentage in the two methods has the less value among the precision and the accuracy values, while in term of accuracy we can find that the logistic regression has a very satisfying performance over KNN classifier. Finally, proposed classifier looks to be the most efficient way to generate a successful anti-spam filter nowadays.

REFERENCES

[1] M. N. Marsono, M. W. El-Kharashi, and F. Gebali, “Binary LNS-based naïve Bayes inference engine for spam control: Noise analysis and FPGA synthesis”, IET Computers & Digital Techniques, 2008

[2] Muhammad N. Marsono, M. Watheq El-Kharashi, Fayez Gebali “Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification” Elsevier Computer Networks, 2009

[3] Yuchun Tang, Sven Krasser, Yuanchen He, Weilai Yang, Dmitri Alperovitch ”Support Vector Machines and Random Forests Modeling for Spam Senders Behavior Analysis” IEEE GLOBECOM, 2008

[4] Guzella, T. S. and Caminhas, W. M. ”A review of machine learning approaches to Spam filtering.” Expert Syst. Appl., 2009

[5] Wu, C. ”Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks” Expert Syst., 2009

[6] Khorsi. “An overview of content-based spam filtering techniques”, Informatica, 2007

[7] Hao Zhang, Alexander C. Berg, Michael Maire, and Jitendra Malic. "SVM-KNN: Discriminative nearest neighbour classification for visual category recognition", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006

[8] Carpinteiro, O. A. S., Lima, I., Assis, J. M. C., de Souza, A. C. Z., Moreira, E. M., & Pinheiro, C. A. M. "A neural model in anti-spam systems.", Lecture notes in computer science.Berlin, Springer, 2006

[9] El-Sayed M. El-Alfy, Radwan E. Abdel-Aal "Using GMDH-based networks for improved spam detection and email feature analysis"Applied Soft Computing, Volume 11, Issue 1, January 2011

[10] Li, K. and Zhong, Z., “Fast statistical spam filter by approximate classifications”, In Proceedings of the Joint international Conference on Measurement and Modeling of Computer Systems. Saint Malo, France, 2006

[11] Cormack, Gordon. Smucker, Mark. Clarke, Charles " Efficient and effective spam filtering and re-ranking for large web datasets" Information Retrieval, Springer Netherlands. January 2011

[12] Almeida,tiago. Almeida, Jurandy.Yamakami, Akebo " Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers" Journal of Internet Services and Applications, Springer London , February 2011

[13] Yoo, S., Yang, Y., Lin, F., and Moon, I. “Mining social networks for personalized email prioritization”. In Proceedings of the 15th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Paris, France), June 28 - July 01, 2009

