ISSN 2278-3091
Volume 8, No.3, May - June 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse73832019.pdf
https://doi.org/10.30534/ijatcse/2019/73832019

A Comparative Analysis of Machine Learning Techniques for Spam Detection

Syed Ishfaq Manzoor[1], Dr Jimmy Singla*
[1]Lovely Professional University Punjab, India, esishfaq@gmail.com
*Lovely Professional University Punjab, India

ABSTRACT

Data Science is an emerging multidisciplinary field which employs algorithms, processes, scientific methods to extract information and insights in various forms which is both structures and unstructured much similar to data mining and prediction analysis. Advertisement and bulk emails, also called as spam, makes an estimate of 62% of the Worldwide internet traffic. Since 1978, when first unwanted mail was sent, technology have advanced but still the detection of spams remains a chronophagous and big budget problem in the field of mathematical sciences. The current study evaluates the effectiveness and efficiency of various machine learning techniques which include K-NN, Decision tree, random forest, Naive Bayes and SVM for spam detection. A data set comprising of 962 emails containing both genuine emails and spams has been used in this study. Some deep learning techniques for classification of spams is also suggested for better performance**.**

Key words: data mining, email, spamming, machine learning, spam detection

1. INTRODUCTION

Email spam may be classified as the any sort of message sent electronically which is undesired to the receiver. Majority of these unsolicited emails are commercial in nature which may also contain hyperlinks which appear to be links of well-known websites but in real lead to phishing web pages owned by hackers and internet fraudsters. The links may also lead to websites hosting malware and ransomwares.

Instant messengers like Facebook and WhatsApp have gained an exponential popularity in recent past however email remains an ascendant medium of communication for individuals and corporations. As per approximations made by Symantec Corporation [1], near about 200 billion messages were sent on daily basis in 2015 through various mail servers. Approximately a common commercial user receives and send at least 42 emails on daily basis. Keeping in view these estimates and approximations, it is easy to conclude why the email is choicest means for fraudsters, hackers and cybercriminals to broadcast to targeted audiences the malicious messages. As per research findings of Nucleus Research [2], US business spent on an average $712 annually on employees for disparaged productivity, astray purchasers, bandwidth consumption and escalating costs of maintenance.

As per Estimates Statista [3] at least 67 percent of incoming business mail traffic is either promotional or unsought bulk mails, known as spams which is lowest since 2003 because of intervention machine learning techniques. In spite of decreasing global spam/non-Spam ratio, there is exponential increase in competition between spammers and spam detection techniques. The problem of getting away with these issues is still persistent and hence need of efficient and effective techniques which classify spams and non-spams is very much needed. The need for automatically classifying spam and non-spam emails by incorporating Deep learning and machine learning techniques is gaining popularity among researchers and academia.

Knowledge Engineering, Deep learning and Machine learning are primary approaches to tackle the spam filtering problems [19, 20]. The prime focus of Knowledge Engineering is on conceiving a knowledge-based system which have predefined criteria known as rules to classify the incoming message into spam or non-spam mail. The prime disadvantage of these methods is that they need to updated and maintained regularly by the users or some third-party vendors. On Contrary deep learning and machine learning approaches don't require such type of predefined rules but previously classified spams and non-spams act as training mails (data set) through which system is trained by deploying a learning algorithm. Thus systems are made intelligent enough for classification of spams. The success ration of these machine learning algorithms varies a lot. Figure 1 shows the Spam/Ham Detection.

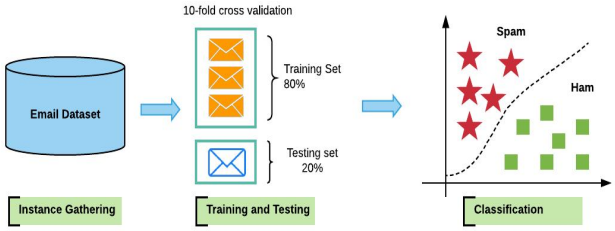


Figure 1: Spam/Ham Detection

For classification of mails into spam and non-spam mails, this study evaluates six Techniques/Algorithms which are as following:

- Support Vector Machines (SVM)
- Decision Trees Algorithms
- Random Decision Forest
- k-Nearest Neighbors (K-NN)
- Adaboost
- Naïve Bayes

2. RELATED WORK

various studies have been carried out in the direction of spam detection and classification since long time. The prime focus of such studies has been spamming detection only. Initially filtering of spam emails was proposed by Sahami et al [9]. The experimented results depict that the classifier employing Bayes Algorithm has better efficiency keeping in view domain specific features in addition uncleaned text of the emails. Still Bayesian Approach is considered as efficient and effective filtering mechanism and is implemented both in clients as well as servers.

Unlike email systems, web is massive and huge which is speeded over geographically scattered nodes [10]. This characteristic feature of web makes it challenging to detect and classify web spams. The authors of [11] where first to formalize and propose a solution for detection of web spams. They proposed TrustRank algorithm to calculate trust score of webpages. Based on ranking of this, the pages with higher scores are considered a trustworthy pages and pages with low score are filtered by search engines as web spams. The authors of [12] proposed as metrics identified as Spam- Mass which was based on link structure for identification and classification of link spamming. The Authors of [13] suggested direct graph model of the web. To detect spam links authors have introduced algorithms based on directed graphs. The authors of [14] proposed both contents based as well as link-based features. They proposed decision tree to classify the web spam. In their research work [15] proposed and implemented semi supervised learning algorithms to boost performance of classifiers with a minimum amount of labelled data samples. Authors of [16] proposed methodologies for spam call detection over IP telephony which was also known as SPIT in VoIP systems. The authors of [17] investigated promotion detectors and video spammers in YouTube. There are a few works for studying spam detection Social Networking Sites and this [18] is one among them. The Authors fetched three data sets of the twitter data. They studied behavior of users, their geographical locations, size of the network, growth patterns of these tweets to classify them as spam and non spams.

A performance comparison of different machine learning classifiers or models which are trained and tested on the basis of training and testing datasets to check the highest accuracy of a model is given so that we can use a model that gives us the accurate and legitimate spams.

3. MACHINE LEARNING TECHNIQUES

Machine Learning and Deep learning are approaches for data analysis and prediction that automate conventional analytical models. It is a sub domain of Artificial Intelligence based on fact that machines can learn from data, machines can identify patterns, extract key features for decision making with a minimal human intervention. The following are popularly used Machine Learning techniques employed for classification and building of systems for auto spam detection:

A. Support Vector Machine

Support Vector Machine (SVM is class of Supervised Learning Algorithms which solves the classification and regression problems much effectively than Knowledge Engineering models. In Support Vector Machines each object is plotted against a n dimensional space where n is number of distinguished features which have been extracted for the classification purpose. The concept of hyperplane is used to separate the two classes as shown in the figure (fig 2).

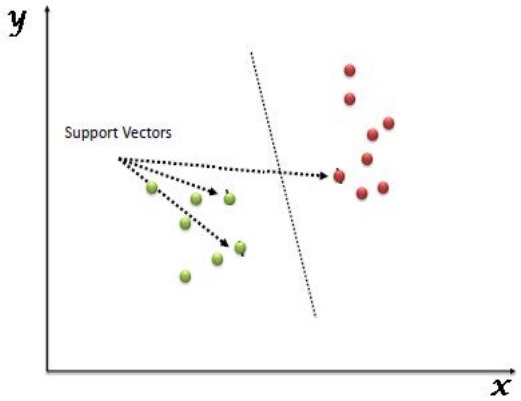


Figure 2: SVM Classification

Vectors in SVM are basically coordinates of individual observations plotted. SVM is among the initial algorithm which segregates two classes (hyper planes/lines) efficiently. Vladmir N. Vapnik and Alexy Ya originally conceptualized the SVM algorithms in their research work [8].

B. Decision Tree Classifier

The classifier of decision tree repeatedly distributes the plot into smaller parts by classifying identifiable lines. Such decision is used in real life scenarios and have significant applications in numerous areas of Predictive analysis and machine learning varying from regression to classification.

.

$$H = -\sum p(x) \log p(x)$$

Where H is entropy

C. Random Forest

Random Decision forests or Random Forests are supervised classification algorithms that ensemble learning methods for classification and regression. As the name depicts in this algorithm many trees together form a forest. Generally, more trees in the forest more is the robustness in the forest. Similarly, in random forest classifiers, higher the number of trees, higher is the accuracy of the classifier. The algorithms deduce the classification labels to form a new document from previously formed decision trees. This algorithm is called Random, because it creates a decision tree by selecting a random set of features from the chosen sample of training data.

D. KNN

K-nearest neighbor (K-NN) classifier is an efficient technique for classification and recognition. In this method, the training data is employed for comparison and not for categorization. Sometimes a term lazy learner is also associated with this classifier.

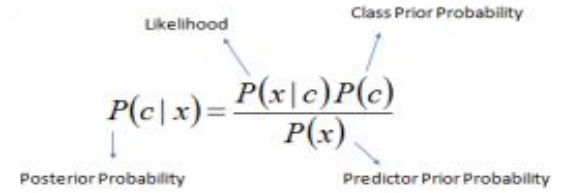
For classification the KNN finds K nearer documents (or pixels /objects, in case of images), uses their profile to find the new weight of the document under observation.

E. Naïve Bayes

The Naive Bayes classification Algorithm rest on Bayes laws which is based on the assumption that there is complete or partial independence among majority of System predictors. In other ways it is assumed that presence of any feature in a system is not affected by other features of the system

Naive Bayes model is easy to rests on and helpful for almost all large data sets. Apart from simplicity of simplicity of Naïve Base method, Naive Bayes classifier outperforms even some modern day sophisticated algorithms.

Posterior probability is calculated as P(c|x) from P(c), P(x) and P(x|c) by methods provided by Bayes Theorem



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

D. Adaboost

AdaBoost, short for Adaptive Boosting is meta algorithm which conjugate various week classifiers into a more effective, efficient and powerful classifier. Week classifiers performs out of sorts but is more desirable than random guessing. A simpler example might be classifying an apple into Grade 1 and Grade 2 category based on its weight and color. We could say an apple with weight of more than 250 gram and red color is grade 1, otherwise grade 2. There is every possibility of misclassification of apples with weight 240 grams and color near to red. The accuracy of such classifiers will start from even at 50 percent.

Adaboost is built on top of such week classifiers and many of such week classifiers form its base and as result on top of many such week classifiers we get a powerful classifier.

4. FEATURE SELECTION

Given a training set of spam emails or messages, a total number of 100 words are chosen through which feature vectors are prepared.

- Spamicity

P(word|Spam) = it gives the probability that a particular word is contained in the spam messages

Pr(word|Ham) = it gives the probability that a particular word is contained in the ham messages.

Spamicity (for a particular word) is defined as:

$$Spamicit(word) = Pr(word|Spam) Pr(word|Spam) + Pr(word|Ham)$$

The relevance proportion of textual messages is considered to come up with estimations of the probabilities. The word specificity in the spam messages is used to come up with the estimation of Pr(word|Spam). The word specificity in the ham messages is used to come up with the estimation of Pr(word|Ham).

- Word Selection

This involves the process of ranking ever word and selecting a total of 100 top ranked words.

Assigning the highest rank to a specific word is intuitively dependent on the spamicity of a word and is far away from (0.5, higher or lower). A closeness to 1 indicates that the message could be a spam and a closeness to 0 indicates that the message is ham.

It has also been seen that the spamicity alone is not adequate enough to serve as the indicator to ham or spam. The words with smaller values for Pr(word|Spam) and Pr(word|Ham) with the spamicity vales even far away from 0.5 do not serve the purpose of being good indicators.

It has therefore been observed to look for the magnitude of the difference:

$$|Pr(word|Spam) - Pr(word|Ham)|$$

Word Selection process:

- The words with $|spamicity - 0.5| < 0.05$ are filtered out.

- A threshold (given as 1% maybe) is used to eliminate the words with rare appearances in the ham and spam messages.
- Calculate |Pr(word|Spam) - Pr(word|Ham)| for all the remaining words and select the top 100 words.

5. EXPERIMENT AND RESULTS

The ML techniques discussed in the previous sections were implemented in python. A dataset with 962 samples that encompassed both legitimate emails and spams was used for training of the models and the testing data set with 260 data samples including both legitimate and spams was used for the comparison of the techniques.

The following table provides the accuracy of each technique.

Table 1: Accuracy

Technique	Accuracy
SVM	97.33
k-NN	92.69
Adaboost	96.15
Decision Tree	92.30
Random Forest	97.30
Naïve Bayes	96.15

A bar graph of the accuracy vs classifier is given in the Fig. 3. below.

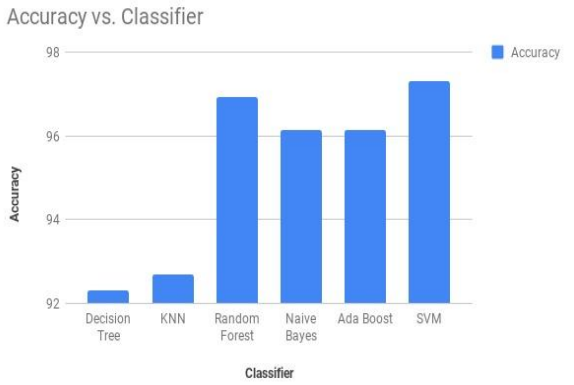


Figure 3: Accuracy Vs Classifier

It is not hard to notice from the results that the random forest approach in addition to SVM is the gold way.

6. CONCLUSION

The paper focused on six machine learning techniques that were used for spam detection on the given data set. It was established that random forest and Support Vector Machines provide better accuracy rates than the remaining approaches. It may be vital to point out that no fine tuning any of those models was done at all and in future works a refined comparison of the same techniques can be provided. Deep learning algorithms like RNN etc can also be implemented in future work

REFERENCES

1. G. Symantec Corporation. (2016). Internet Security Threat Report (Vol. 21).
2. Nucleus Research. (2007). Spam costing US Businesses $712 Per Employee.
3. Statista. (2017). Global spam email traffic share 2014-2017.
4. Tin Kam Ho. (1995). Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition, 1, 278–282. https://doi.org/10.1109/ICDAR.1995.598994
5. Stephen Marsland. (2014). Machine Learning: An Algorithmic Perspective (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/b17476
6. Breiman, L. (2001). Random Forests. Machine Learning, 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324
7. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. https://doi.org/10.1007/978-1-4614-7138-7
8. Vapnik, V., & Chervonenkis, A. (1964). A note on one class of perceptrons. Automation and Remote Control, 25.
9. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A bayesian approach to filtering junk e-mail. In: AAAI Workshop on Learning for Text Categorization (1998)
10. Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., Raghavan, S.: Searching theweb. ACM Trans. Internet Technol. 1(1), 2–43 (2001) https://doi.org/10.1145/383034.383035
11. Gy¨ ongyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with trustrank. In: Proceedings of the Thirtieth international conference on Very large data bases, p. 576–587 (2004) https://doi.org/10.1016/B978-012088469-8.50052-8
12. Gyongyi, Z., Berkhin, P., Garcia-Molina, H., Pedersen, J.: Link spam detection based on mass estimation. In: VLDB 2006: Proceedings of the 32nd international conference on Very large data bases, pp. 439–450 (2006)
13. Zhou, D., Burges, C.J.C., Tao, T.: Transductive link spam detection. In: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web, pp. 21–28 (2007) https://doi.org/10.1145/1244408.1244413
14. Castillo, C., Donato, D., Gionis, A., Murdock, V., Silvestri, F.: Know your neighbors: web spam detection using the web topology. In: Proceedings of the 30th annual international ACM SIGIR conference, pp. 423–430 (2007) https://doi.org/10.1145/1277741.1277814

15. Geng, G.G., Li, Q., Zhang, X.: Link based small sample learning for web spam detection. In: Proceedings of the 18th international conference on World wide web, pp. 1185–1186 (2009) https://doi.org/10.1145/1526709.1526920
16. Wu, Y.-S., Bagchi, S., Singh, N., Wita, R.: Spam detection in voice-over-ip calls through semi-supervised clustering. In: Proceedings of the 2009 Dependable Systems Networks, pp. 307–316 (2009) https://doi.org/10.1109/DSN.2009.5270323
17. Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., Gon¸calves, M.: Detecting spammers and content promoters in online video social networks. In: Proceedings of the 32nd international ACM SIGIR conference, pp. 620–627 (2009) https://doi.org/10.1145/1571941.1572047
18. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: WOSP 2008: Proceedings of the first workshop on Online social networks, pp. 19–24 (2008) https://doi.org/10.1145/1397735.1397741
19. B.Manoj, K.V.K.Sasikanth, M.V.Subbarao and V Jyothi Prakash, Analysis of Data Science with the use of Big Data, IJATCSE, volume 7 no 6, pp- 87-90, 2018 https://doi.org/10.30534/ijatcse/2018/02762018
20. Apoorva Deshpande, Ramnaresh Sharma, Multilevel Ensembler Classifier using Normalized Feature Based Intrusion Detection System, IJATCSE, volume 7 no 5, pp 72-76, 2018. https://doi.org/10.30534/ijatcse/2018/02752018