iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 Issue: VI Month of publication: June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6159

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

E-Mail Spam Detection using Machine Learning and Deep Learning

Shivam Pandey[1], Ashish Taralekar[2], Ruchi Yadav[3], Shreyas Deshmukh[4], Prof. Shubhangi Suryavanshi[5]
[1, 2, 3, 4]Student, [5]Guide , Department of Computer Engineering, G.H. Raisoni Institute of Engineering & Technology, Wagholi, Pune – 412207

Abstract: Here we present an inclusive review of recent and successful content-based e-mail spam filtering techniques. Our focus is primarily on machine learning-based spam filters and variants that are inspired by them. We report on related ideas, techniques, major efforts and cutting-edge art in the field. Preliminary interpretation of prior work shows the basics of e-mail spam filtering and feature engineering. In this we conclude by studying techniques, methods, evaluation criteria, and explore promising apprehensions of the latest developments and suggest future lines of investigation.
Keywords: SVM Classifier, Spam Email Classification, Data Mining, Machine Learning and Deep Learning, etc

I. INTRODUCTION

In present times the commercial or bulk e-mails have become a really major problem. Spam nowadays is a waste of storage space, time and bandwidth for communication. From many years the problem caused by spam or fraud mails is increasing. In recent studies, 77% of all mail is spam that comes around a value of 15 billion emails per day and costs Internet users about $ 300 million per year.

Today for email filtering, knowledge Engineering and Machine Learning are two most successful approaches. In knowledge engineering approach the hard and fast rule is specifying a set of principles according to which email is classified as spam or ham. Application of this method, doesn't shows any promising results because the rules should be necessary. Constantly updating the rules and methods just causes waste of time and requires more maintenance. As compared to knowledge Engineering, Machine learning is more appropriate approach.

It does not have to specify any rules. A set of pre-classified e-mail messages is used here in place of set of rules. Machine learning approaches have a wide range of Importance and a lot of algorithms can be used for e-mail filtering and classification. These include Support Vector Machine, Naïve Bayes.

II. PRELIMINARY AND PROBLEM STATEMENT

A Spamming is one of the major and common attacks that accumulate a large number of compromised machines by sending unwanted messages, viruses, and phishing through email. We have chosen this project because now there are many people who are trying to fool you just by sending you fake e-mails, as if you have won 1000 dollars, deposit this amount in your account as soon as you open this link. Once done, they will track you and try to hack your information. Sometimes relevant e-mail is considered spam email.

Unwanted email is harassing Internet consumers in ways such as:

1) Important email messages were missed and / or delayed.
2) Consumers seek ISP's frequent email delivery changes all the time.
3) Internet performance and bandwidth loss.
4) Millions of compromised computers.
5) Loss of billions of dollars worldwide.
6) Increase in several viruses and Trojan horses.

III.PROPOSED SYSTEM

A. Machine Learning

Spam filtering, from the aspect of machine learning, is essentially a classification problem in which we aim to classify an email as spam or ham which is dependent on its feature. For instance, (x, y) can be a data point where x is a dimensional vector containing the features and y is either 1 or 0, which indicates spam or ham. Systems with machine learning can be taught or trained to classify emails.

©IJRASET: All Rights are Reserved

1) *Support Vector Machine:* Support Vector Machines with associated learning algorithms analyses data for classification. The SVM algorithm for training constructs a model which allocates the new examples into one of the categories. In an SVM model examples are represented as points in n- dimensional space which are mapped so that the points of the different categories are separated by a gap that should be as broad as possible. Then the untrained examples are mapped in that space and a decision is made, that is, to which category does it belongs to depending upon the side of the plane they fall.

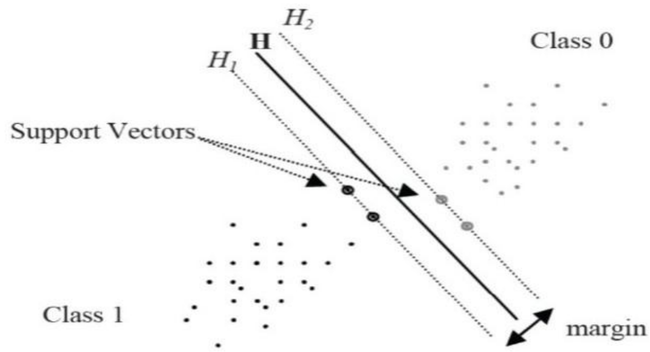


Fig -1: A hyperplane separating two classes

2) *Naïve Bayes:* The roots of the Naïve Bayesian classifier lie in the Bayes Theorem.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Bayes Theorem basically describes how much we should modify the probability so that our hypothesis (H) transpires, given some novel evidence (e). This paper determines the probability that an email is spam, given the evidence of the email's feature values F_1, F_2….,F_n. These features are just a Boolean value (0 or 1) dependent on whether the feature is present in the email or not. Then P (Spam| features) to P (Ham| features) are determined and then decided which is more likely.

B. Deep Learning

In this paper, we exploit a deep neural network for E-mail Spam Detection using TensorFlow. We build the neural network model which contains recurrent neural networks and LSTM (Long Short-Term Memory) which automatically extracts the features avoiding the overhead of exclusively extracting the features. We are training and testing the model on our self-designed dataset. Results on our dataset show that the neural model achieves significantly better accuracies compared to the previous studies done on E-mail Spam Detection using linguistic approach, demonstrating the advantage of the automatically extracted neural features.

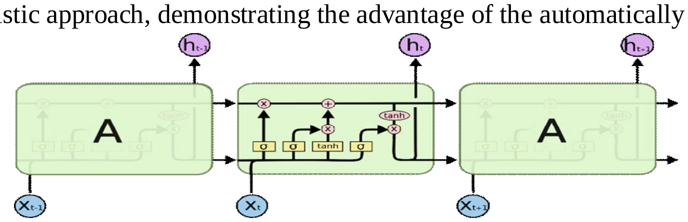


Fig -2: LSTM Networks

1) *Data Collection*

a) *Extraction:* We downloaded the data from Kaggle which consist of more than 900 spam emails and ham emails. We divided our complete file into three major parts spam training data, ham training data, and test data. This data set consist of many URL's, symbols and keywords as well which were not important to our model and thus were removed from our file during preprocessing.

b) *Preprocessing:* The input data for our RNN model is a single text file containing training and testing data in alphanumeric format. This includes 3 blocks of data, two training blocks with spam and ham (meaning no spam) examples and one block of mixed spam / ham to test our solution. The block is divided by header lines. Each data line begins with a label (spam or ham) after which the text is evaluated. First, we will separate the training lines from the test lines, preserving the original line format. We will use '#' test data for this separation. We will also be altering training and testing data. Second, we will divide the two blocks into labels and data. We will remove some formatting information, but keep the alphanumeric format for now.

©IJRASET: All Rights Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

2) *Sequences and Tokenizers:* We use the tokenizer class from the pre-processing package to convert our text to vectors. The tokenizer method is initialized using our training data (text part only). This will convert the text into a dictionary of words. Then we will convert this list of indices to a binary NumPy matrix. Matrix columns represent words in text data, rows represent text lines. We will create a second NumPy matrix for the test data. In this case we will only use terminology from training data, as our model will be trained on it. So, the second matrix has the same column created from training data and binary flags created from the test set.

IV. RESULT AND ANALYSIS

A. Machine Learning

Here we compare the accuracy and performance of SVM with Naïve Bayes Classifier for the same set of data. Following images consist of the factors that are being compared.

1) *SVM Classifier*

```
Accuracy score for SVM is: 0.9752559726962458
Confusion Matrix for SVM is:
 [[876   7]
 [ 22 267]]
Classification report for SVM  is:
              precision    recall  f1-score   support

           0       0.98      0.99      0.98       883
           1       0.97      0.92      0.95       289

    accuracy                           0.98      1172
   macro avg       0.97      0.96      0.97      1172
weighted avg       0.98      0.98      0.98      1172

F1 Measure for SVM is: 0.9484902309058615
saved
```

Fig -3: Metrics for SVM

The above Fig. displays the factors like confusion matrix, classification report and f1 measure for SVM Classifier.

2) *Naïve Bayes Classifier*

```
Accuracy score for MNB is: 0.9803754266211604
Confusion Matrix for MNB is:
 [[870  13]
 [ 10 279]]
Classification report for MNB is:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       883
           1       0.96      0.97      0.96       289

    accuracy                           0.98      1172
   macro avg       0.97      0.98      0.97      1172
weighted avg       0.98      0.98      0.98      1172

F1 Measure for MNB is: 0.9604130808950087
```

Fig -4: Metrics for MNB

3) *Predicted Output*

```
In [3]: runfile('E:/BE PROJECT_Final/project-material 2020/SVM_detector.py', wdir='E:/BE PROJECT_Final/project-material 2020')
Which algorithm do you want to use for the detection

1.SVM
2.MNB

Enter your choice
```

```
>Subject: letter to academics
    ...: vince ,
    ...: ?
    ...: hi , sorry to bother you , but was wondering if you ' ve heard anything about
    ...: seeking approval to use the text in the letter i sent last week ? ? we would
    ...: like to send this out early next week so if you know of someone who i should
    ...: contact within corporate communications , please let me know .
    ...: ?
    ...: thanks , and hope all is well .
    ...: ?
    ...: julie
Not Spam
```

©IJRASET: All Rights are Reserved

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

B. Deep Learning

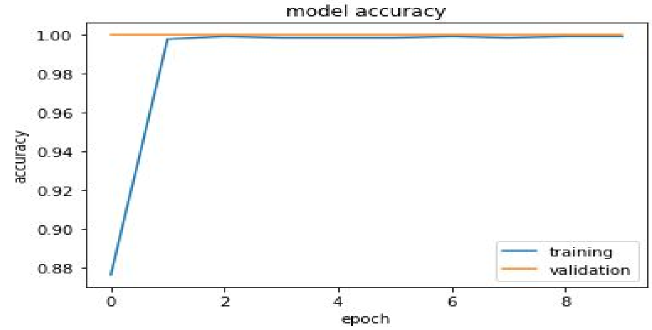


Fig-5: Model Accuracy

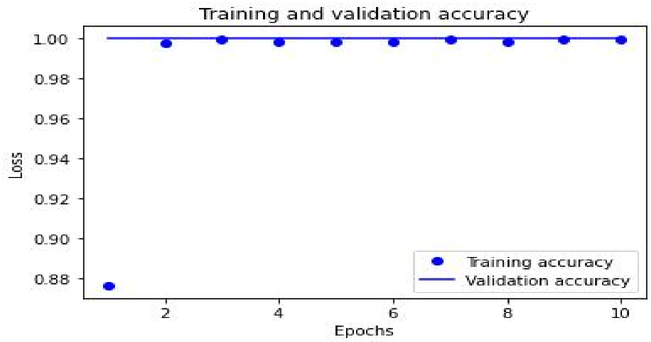


Fig-6: Training and Validation Accuracy

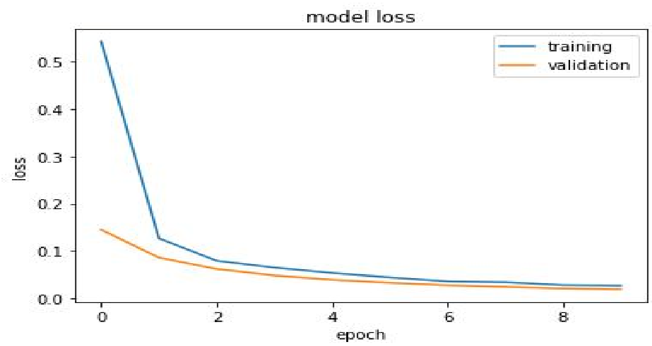


Fig-7: Model Loss

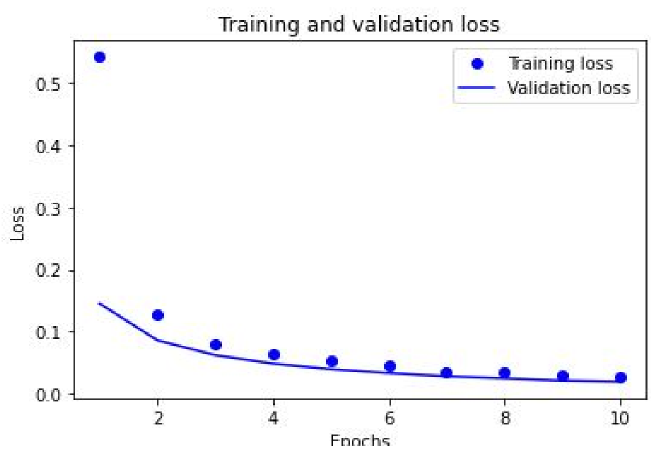


Fig-8: Training and Validation Loss

©IJRASET: All Rights are Reserved

V. FUTURE SCOPE

However, the experiment has made efforts towards solving the problem of spam e-mail. Proposed solutions using legislative, behavioural and technical measures are not a complete solution. The problem of spam e-mail and anti-spam solutions is game like cat and mouse, every day spammers will come up with new techniques Send spam e-mail. This work has given possible directions for classification. Spam e-mail Future efforts will be extended to:

A. Obtaining accurate classification, zero percent (0%) with abortion of ham E-mail as spam and spam as e-mail ham.
B. Many Efforts will be implemented to block phishing e-mail, which carries phishing Attacks and now days which is a matter of concern.
C. Also, work can be extended to keep it away from the Denial of service attack (DoS). Now which has emerged in distributed fashion, is called distributed Denial of Service Attack (DoS).

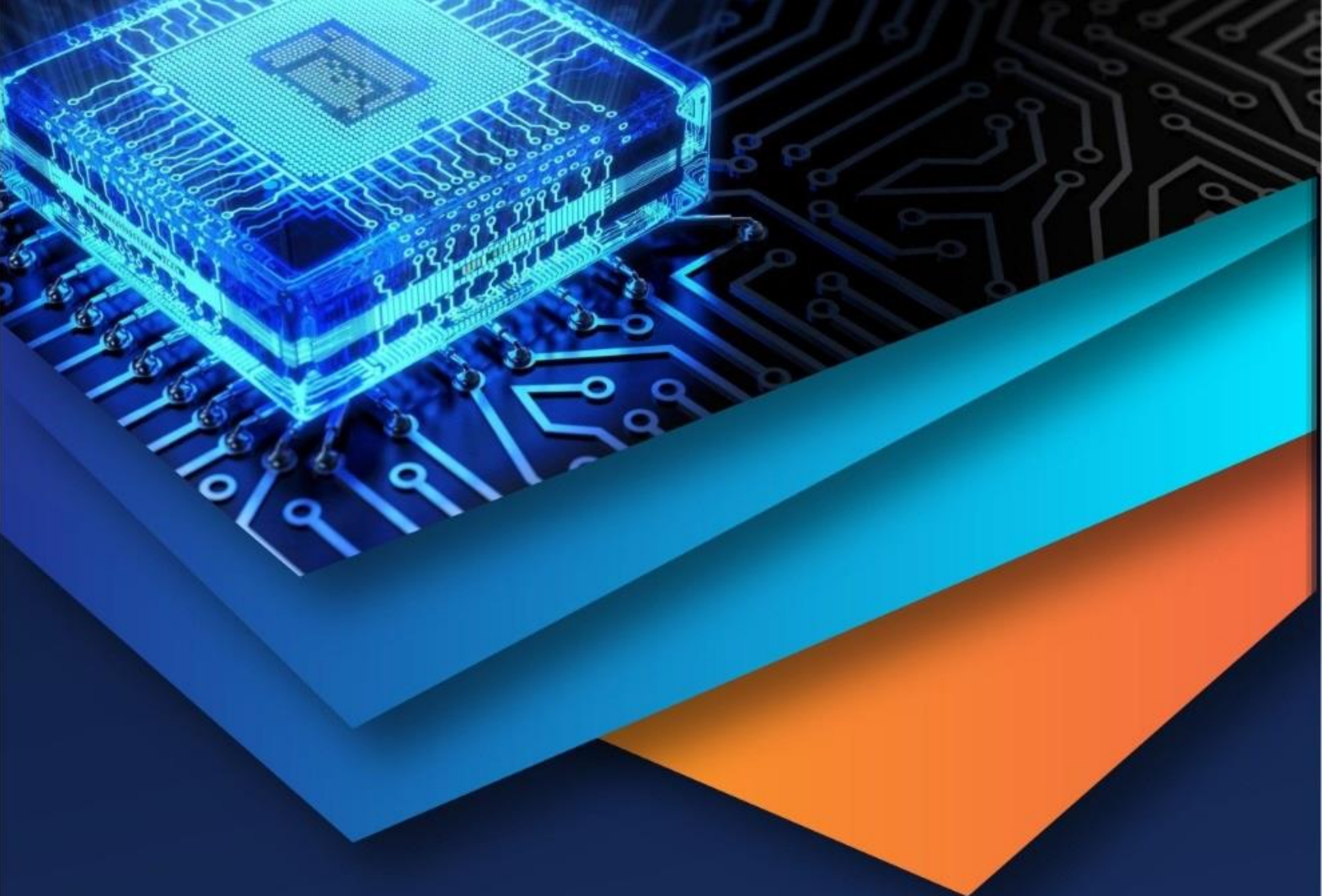
VI. CONCLUSION

In this study, we reviewed the general application in the field of machine learning approach and spam filtering. A review of the state-of-the-art algorithm has been implemented to classify the message as either spam or ham. Efforts made by various researchers to solve the problem of spam through the use of machine learning classifiers were discussed. The development of spam messages was investigated over the years to avoid filters. The basic structure of the email spam filter and the processes involved in filtering spam emails were noted. The paper surveyed some of the publicly available datasets and performance metrics that can be used to measure the effectiveness of any spam filter. The challenges of machine learning algorithms in efficiently handling the threat of spam were pointed out and a comparative study of machine learning techniques available in the literature. We also revealed some open research problems related to spam filters. In general, the amount and amount of literature we reviewed suggests that significant progress has been made and will still be made in this area. After discussing open problems in spam filtering, further research needs to be done to increase the effectiveness of spam filters. It will develop spam filters to continue an active research area for academics and industry practitioners researching machine learning techniques for effective spamming. Our hope is that research students will use this paper as a spring board to conduct qualitative research in spam filtering using machine learning, deep learning, and deep adversarial learning algorithms.

REFERENCES

[1] Abduelbaset M. However, Tarik Rashed, Ali S. Elbekaie, and Husien A. Alhammi, “An Anti-Spam System Using Artificial Neural Networks And Genetic Algorithms” (A Neural Model In Anti Spam).
[2] Er. Seema Rani, Er. Sugandha Sharma, “Survey on E-mail Spam Detection Using NLP”, International Journal of Advanced Research in Computer Science and Software Engineering, India, Volume 4, Issue 5, May 2014.
[3] Masurah Mohamad, Khairulliza Ahmad Salleh, “Independent Feature Selection as Spam-Filtering Technique: An Evaluation of Neural Network”, Malaysia.
[4] El-Sayed M. El-Alfy, “Learning Methods For Spam Filtering”, College of Computer Sciences and Engineering King Fahd University of Petroleum and Minerals, Saudi Arabia.
[5] Upasna Attri & Harpreet Kaur, “Comparative Study of Gaussian and Nearest Mean Classifiers for Filtering Spam E-mails”, Global Journal of Computer Science and Technology Network, Web & Security, USA, Volume 12 Issue 11 Version June 2012.
[6] Alia Taha Sabri, Adel Hamdan Mohammads, Bassam Al-Shargabi, Maher Abu Hamdeh, “Developing New Continuous Learning Approach for Spam Detection using Artificial Neural Network (CLA_ANN)”, European Journal of Scientific Research, ISSN 1450-216X Vol.42 No.3 (2010), pp.511-521.
[7] Enrique Puertas Sanz, José María Gómez Hidalgo,José Carlos Cortizo Pérez, “Email Spam Filtering”, Universidad Europea de Madrid Villaviciosa de Odón, 28670 Madrid, SPAIN.
[8] Ravinder Kamboj, “A rule based approach for spam detection” ,Computer Science and Engineering Department, Thapar University, India, July 2010.
[9] Vandana Jaswal, Nidhi Sood, “Spam Detection System Using Hidden Markov Model”, International Journal of Advanced Research in Computer Science and Software Engineering, India, Volume 3, Issue 7, July 2013.
[10] Sahil Puri, Dishant Gosain, Mehak Ahuja, Ishita Kathuria, Nishtha Jatana, “Comparison and Analysis of Spam Detection.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)