ISSN 2347 - 3983
Volume 8. No. 5, May 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter116852020.pdf
https://doi.org/10.30534/ijeter/2020/116852020

Improving the Accuracy of Spam Message Filtering using Hybrid CNN Classification

Aditi P. Marathe[1], Avinash J. Agrawal[2]
[1]M.Tech Student, Department of Computer Science, RCOEM, Nagpur, India, aditipmarathe@gmail.com
[2]Associate Professor, Department of Computer Science, RCOEM, Nagpur, India, agrawalaj@rknec.edu

ABSTRACT

Spam messages are growing day by day due to the invention of low-cost messaging and emailing solutions. Due to this, the identification of genuine messages from the spammy ones requires a lot of learning. This learning includes training the system for spam messages, and then training another system for non-spam or genuine messages. Once these systems are trained, then a probabilistic classifier is needed, which can find out the probability of the message to either be spam or genuine. Such a network is called as two-stage convolutional neural network. In this paper, we have designed a two-stage convolutional neural network, that first trains one network with spam messages, and then trains another network with non-spam messages. These stages are cascaded, and the outputs of each stage is given to a decision unit. The unit evaluates the probabilities of spam and non-spam messages, and finally classifies the input text into either spam or non-spam. The proposed system is tested on the standard UCI spam text dataset, and it has achieved more than 90% accuracy for classification of spam messages.

Key words: Recommendation, classification, clustering, pre-processing, accuracy.

1. INTRODUCTION

A typical machine learning-based text Spam message filtering system requires a lot of language processing, and filtering units. Such a system is described in figure 1. From the figure we can observe that the spam filtering system first performs pre-processing of the input messages. This includes stemming, stripping, parts-of-speech tagging, chunking, etc. All these operations results in a pre-processed text which can be further used for complex computations.

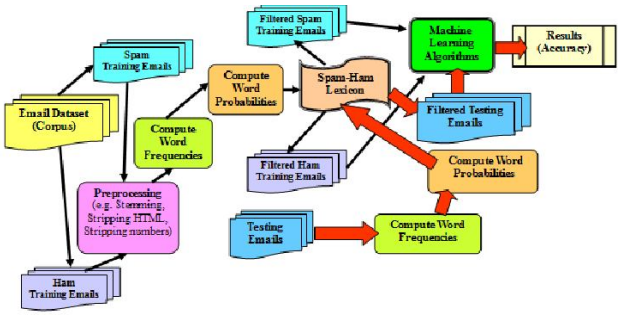

Figure 1: Typical spam filtering system

These computations include counting the work frequencies, evaluation of work probabilities, lexicon comparison, etc. All these operations helps in finding out a preliminary level of spam messages, which are then removed from the system. These operations also assist in finding out the feature sets for the input documents. These feature sets must have differentiation between normal and spam messages. Finally, these features are given to a machine learning classifier like CNN, and the resulting accuracy is evaluated. This accuracy is further improved by adding more messages to the training set, or increasing the number of iterations/epochs, increasing the number of neurons per layer, etc. We also propose the use of a similar structure, but we have modified the CNN into a two-stage network, where one CNN each is trained for spam and ham (non-spam) messages. The next section describes various spam detection systems and their nuances, followed by our approach, and finally the result evaluation of the proposed approach on the given dataset. We conclude the paper with some interesting observations about the proposed CNN model, and some further analytics which can be done in order to extend this work.

2. LITERATURE REVIEW

For discovering spam messages a few methodologies are recommended, for instance the standard based methodology is utilized by making rules to order the approaching messages. It is known as the immediate methodology. It doesn't require any preparation stage. Rules spread various

dangers, dubious arrangement and powerless birthplace inclined to sending spam implies the sender is affirmed as open transfer. While utilizing this methodology we must be cautious since rules created by us can lead the approaching messages to misclassification. There is danger of bogus negative and bogus positive. Learning-based methodology manages preparing of Spam channel. A huge arrangement of ham messages and spam messages is utilized to prepare the spam channel. In preparing channel peruses tokens from messages and change the estimations of tokens/words in the database as per their classification whether they are from spam message or ham message. So as to arrive at most extreme precision and speculation abilities classifiers must concentrate just relevant data from the preparation information. A robotized procedure to amass related records is known as Clustering. Based on comparable qualities for traits related records are gathered. In grouping examination approach, it isn't essential that the end-client/investigator may determine heretofore how records should be connected together. The target of the examination is, truth be told, for the most part to find bunches or portions and afterward assess the qualities and traits that characterize the groups Classification is indistinguishable from grouping as it likewise arranges client records into particular fragments called classes. In any case, not at all like bunching an arrangement investigation necessitates that the end-client/examiner know early how classes are characterized. Robotized spam sifting can be considered as a basic occurrence of report characterization. In report arrangement issues, we have two arrangements of records. The principal report set has a predefined class and is known as the preparation set of records. This archive set is utilized by the classifier to learn designs in the information. The subsequent archives set don't have the class marks with it and is utilized for the testing reason. These archives set establish all models from this present reality which will be given as contribution to the characterization calculation to group later on. The issue of spam location is fundamentally the equivalent with as that of report arrangement with two classes for example Spam and Legitimate. The activity of our sifting procedure is to accept messages as sources of info and attempts to find out about examples that will speaks to various classes. When the learning is done, at that point given an obscure occurrence of message it ought to have the option to sift through spam with high precision. In [1] analysts had accomplished an achievement work to recognize cell phone spam and evaluated a few Bayesian based classifiers [2]. In this work, the initial two notable SMS spam datasets, to be specific, the Spanish and English test databases were proposed by the creators. Various message depiction techniques and AI approaches were tried by the creators on those two datasets. They concocted the end that Bayesian separating methods can be enough utilized to order SMS spam. The specialists assessed that even substance-based spam sifting cab be utilized for short instant messages which happens in three assorted points of view: SMS, blog remarks, and email synopsis data [3]. Creators reasoned that SMS are confined to have inadequate words for the best possible help of words or word bigram-based spam classifiers. In this way, the channel's effectiveness was improved by growing the arrangement of highlights to incorporate symmetrical meager word bigrams and furthermore to incorporate character bigrams and trigrams. They executed DMC, a pressure model-based classifier, which doesn't depend on express featurization and performed well on short messages and message sections.

In [4] scientists investigated the productivity of sieving message spam on autonomous phones utilizing Text Classification approaches [5]. On a free versatile different preparing were done identified with preparing, separating, and refreshing. Their built-up results show that the anticipated model was effective in refining messages hams and spam with sensible effectiveness, less capacity utilization, and fitting preparing time without taking the assistance from a machine. In [6] scientists gave a system based online location strategy for the distinguishing proof of SMS spams crusade by taking the count of number of messages which were sent in single system over a little timeframe and convey comparative kind of information [7]. The methodology given by them included Bloom channels to keep a provisional tally of message content events. In [8] analysts have taken a shot at a bunching investigate a SMS corpus [9]. To get to the conduct of SMS spam, they accumulated 1353 spam messages and attempted to utilize it as the dataset which grasped of no deception. They applied k-way ghastly bunching with symmetrical instatement. By applying ghastly bunching all alone aggregated dataset scarcely any groups were created which were ten in check with their connected top 8 terms and an assumed explanation.

In [10] scientists exhibited the points of interest of another true, open and non-encoded SMS spam aggregation which comprises of greatest number of messages [11]. It is created by 4,827 versatile ham messages and 747 portable spams. Besides, the creators played out a few set up AI calculations on their dataset and they arrived at the resolution that as indicated by them SVM is a superior methodology for advance assessment as it accomplished great precision. In [12] specialists applied diverse AI calculations to SMS spam grouping issue, contrast their presentation with gain understanding and further investigate the issue, and plan an application dependent on one of these calculations that can channel SMS spams with high precision [13]. They utilized a database of 5574 instant messages.

3. PROPOSED SPAM DETECTION FRAMEWORK

The proposed spam message/email detection framework consists of the following blocks,

- Data pre-processing block to remove any stop words, and find out action words in the text using language processing.

- Feature extraction engine that evaluates the different N-gram based features for the action words.
- Multi-layer CNN classification block, which uses a 2-stage probabilistic classifier to find out the final class of the input sentence.

The block diagram of the proposed system can be observed from the following figure 2.

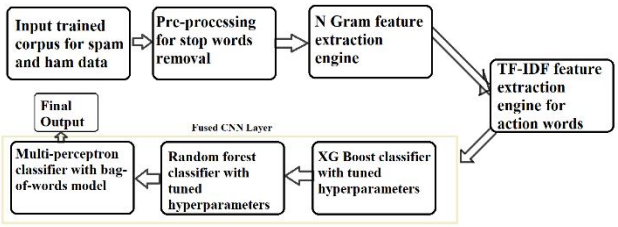

Figure 2: Proposed system block diagram

From the block diagram, we can observe that the input corpus is first given to the pre-processing unit, where first the text is pre-processed to remove any stop words, and then action words are found from the text. This pre-processing can be done using the natural language processing library of your development environment, in our case we have used NLTK in Python. This step removes any kind of stop words from the input document, and provides only the action words at the output. The advantages of using this step are two-fold,

1. It reduces the length of the input document considerably, thereby requiring lesser number of cycles for processing.
2. It removes the stop words, thereby improving the efficiency of feature extraction.

Once the stop words are removed, then a N-gram based feature extraction block is used to find out the co-existence of different words & group of words with each other. This approach allows the system to find out the number of times one word is co-occurring with other words or group-of-words. Thereby, evaluating the document patterns. Once this N-gram is evaluated, then each document is given to a term-frequency & inverse-document-frequency evaluation unit. Wherein the probability of occurrence of each of the N-grams in the document is evaluated. This feature set helps in converting the text document into numerical representation, thereby facilitating the classification process.

Our hybrid convolutional neural network model is a combination of 3 different classifiers,

1. XG Boost classifier to boost the tuned hyper-parameters.
2. Random forest classifier to find out the best features for classification.
3. Multi-perceptron classifier to find out the final class of the input text.

The XG boost classifier is trained for both spam and ham messages, and it helps in tuning the feature vectors (also called as hyper parameters). The feature vectors are tuned so that the overall distinguishability between them increases across spam and ham classes. This tuning helps in optimizing the feature vector selection. Extended Gradient Boosting (or XGBoost) is the best choice for this purpose, because it boosts the gradients (differences) between the feature vectors. The random forest (RF) classifier is introduced after XG Boost to further facilitate feature selection by selecting random features from the XGBoost's output. These random feature sets are then given to a multi-perceptron classifier. This classifier is a flat layered neural network that classifies the input data into spam or ham, and based on the probability of spam and ham the final output is obtained. The proposed system generates very promising results, which are evaluated in the next section. We evaluated these results on the UCI (University of California) datasets for spam and ham messages, which has rapidly become a standard set for classification of spam and non-spam messages.

4. RESULTS AND ANALYSIS

The UCI corpus has been gathered from free or free for investigate sources at the Internet. It is an assortment of 425 SMS spam messages was physically extricated from the Grumbletext Web webpage. This is a UK discussion wherein PDA clients make open cases about SMS spam messages, the vast majority of them without announcing the very spam message got. The recognizable proof of the content of spam messages in the cases is a hard and tedious assignment, and it included cautiously filtering several site pages. A subset of 3,375 SMS haphazardly picked ham messages of the NUS SMS Corpus (NSC), which is a dataset of around 10,000 genuine messages gathered for look into at the Department of Computer Science at the National University of Singapore. The messages to a great extent begin from Singaporeans and for the most part from understudies going to the University. These messages were gathered from volunteers who were made mindful that their commitments would have been made openly accessible. A rundown of 450 SMS ham messages gathered from Caroline Tag's PhD Thesis. At long last, they have consolidated the SMS Spam Corpus v.0.1 Big. It has 1,002 SMS ham messages and 322 spam messages. We evaluated the accuracy of different classifiers on the given dataset across different size of input training sets. We also evaluated the delay of execution for the different algorithms and evaluated their performance on both parameters. It was found that the proposed classifier performs best in terms of accuracy, followed by the XGBoost classifier. But the XGBoost classifier requires a more delay when compared to the other classifiers. The delay of both the proposed classifier

and the XGBoost classifier is almost same. The following accuracy table 1 is obtained for the different classifiers,

Table 1: Accuracy comparison

Number of texts	Accuracy (%) Random Forest	Accuracy (%) XGB	Accuracy (%) Naïve Bayes	Accuracy (%) Proposed
100	80	85	82.5	87
200	82	87	84.5	89
500	85	90	87.5	90
1000	86	91	88.5	91
2000	87	93	90	93
3000	89	94	91.5	95
5000	90	95	92.5	98

A similar comparison for delay is observed in the following table 2

Table 2: Delay comparison

Number of texts	Delay (ms) Random Forest	Delay (ms) XGB	Delay (ms) Naïve Bayes	Delay (ms) Proposed
100	0.80	25.00	8.60	17.20
200	0.90	39.00	13.30	26.60
500	1.10	59.00	20.03	40.07
1000	1.20	62.00	21.07	42.13
2000	1.30	65.00	22.10	44.20
3000	1.35	70.00	23.78	47.57
5000	1.40	75.00	25.47	50.93

These results were also represented in the form of graphs as shown in the following figure 3.

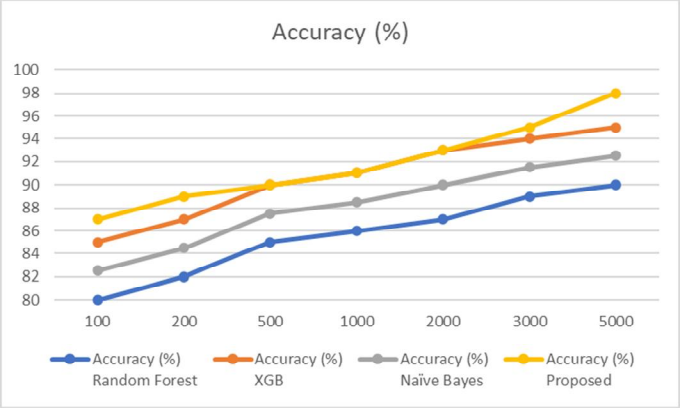

Figure 3: Accuracy comparison

A similar plot was obtained for the delay of execution for these algorithms, and it can be shown in the figure 5. From these figures it is clear that the best accuracy to delay ratio can be obtained from the proposed classifier. The main reason for reduction in delay is the usage of XGBoost for feature differentiation rather than complete classification. Moreover, the delay of the other algorithms is very low when compared to the proposed algorithm, but this can be tackled by using high performance computation techniques like machine learning and AI that are focused on delay optimization.

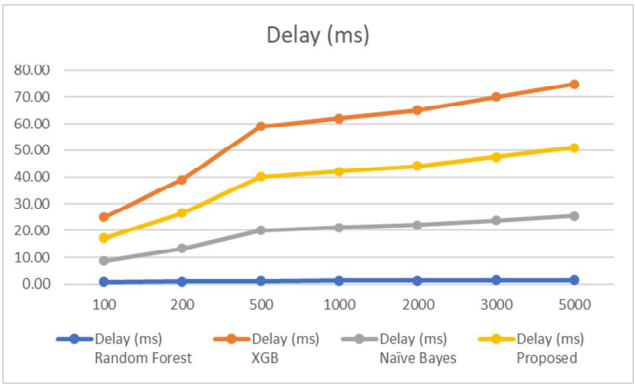

Figure 4: Delay comparison

Thus, we can observe that the proposed algorithm gives the best performance in terms of overall accuracy of classification of spam messages.

5. CONCLUSION AND FUTURE SCOPE

Due to a complex structure, the overall delay of operation of the proposed algorithm is more than most of the simplistic algorithms like Naïve Bayes, Random forest, etc. But, due to this complexity there is a massive improvement in the overall accuracy of spam messages classification. There is an improvement of more than 20% in terms of classification accuracy, which is a huge factor while deciding the usage of any spam detection system. The accuracy can be further improved by adding more training data to the corpus. In future, we would recommend the researchers to work on delay focused AI algorithms like delay aware Genetic Algorithm, Q-Learning algorithms, etc. Moreover, researchers can also focus on improving the quality of Spam detection using blockchain-based technologies.

REFERENCES

[1] Gauri Jain, Manisha Sharma and Basant Agarwal**, Spam detection in social media using convolutional and long short term memory neural network**, *Ann Math ArtifIntell 85*, pp. 21–44, 2019. https://doi.org/10.1007/s10472-018-9612-z

[2] S. Venkatraman, B. Surendiran and P. Arun Raj Kumar, **Spam e-mail classification for the Internet of Things environment using semantic similarity approach**, *J Supercomput 76*, pp. 756–776, 2020. https://doi.org/10.1007/s11227-019-02913-7

[3] Kanaris I., Kanaris K., Stamatatos E. (2006) **Spam Detection Using Character N-Grams**, *Advances in Artificial Intelligence*. SETN 2006. Lecture Notes in Computer Science, vol 3955. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11752912_12

[4] M. Z Asghar, A. Ullah, S. Ahmad and A. Khan, **Opinion spam detection framework using hybrid classification**

scheme, *Soft Computing volume 24*, pp. 3475–3498, 2020.
https://doi.org/10.1007/s00500-019-04107-y

[5] T. Gangavarapu, C. D. Jaidhar and B. Chanduka, **Applicability of machine learning in spam and phishing email filtering: review and approaches**, *Artificial Intelligence Review*, 2020.
https://doi.org/10.1007/s10462-020-09814-9

[6] O. Papapetrou, W. Siberski and W. Nejdl, **Cardinality estimation and dynamic length adaptation for Bloom filters**, *Distrib Parallel Databases 28*, pp. 119–156, 2010.

[7] Komal Dhingra and Sumit Kr Yadav, **Spam analysis of big reviews dataset using Fuzzy Ranking Evaluation Algorithm and Hadoop**, *International Journal of Machine Learning and Cybernetics volume 10*, pp. 2143–2162, 2019.
https://doi.org/10.1007/s13042-017-0768-3

[8] Ankit Kumar Jain, Diksha Goel, Sanjli Agarwal, Yukta Singh and Gaurav Bajaj, **Predicting Spam Messages Using Back Propagation Neural Network**, *Wireless Personal Communications volume 110*, pp. 403–422, 2020.

[9] Kołcz A. (2011) **Text Mining for Spam Filtering**. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA.

[10] H. Afzal, K. Mehmood, **Spam filtering of bi-lingual tweets using machine learning**, *18th International Conference on Advanced Communication Technology* (ICACT), 2016.
https://doi.org/10.1109/ICACT.2016.7423529

[11] J. Qiu, W. Luo, L. Pan, Y. Tai, J. Zhang and Y. Xiang, **Predicting the Impact of Android Malicious Samples via Machine Learning**, *in IEEE Access*, vol. 7, pp. 66304-66316, 2019.
https://doi.org/10.1109/ACCESS.2019.2914311

[12] Kozik, Rafal, Luckner, Marcin**, Practical Web Spam Lifelong Machine Learning System with Automatic Adjustment to Current Lifecycle Phase**, Security and Communication Networks, Hindawi.

[13] S. Kumar, X. Gao, I. Welch and M. Mansoori, **A Machine Learning Based Web Spam Filtering Approach**, *IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, Crans-Montana, pp. 973-980, 2016.
https://doi.org/10.1109/AINA.2016.177