IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Review of Computer Engineering Studies
Vol. 7, No. 4, December, 2020, pp. 91-95
Journal homepage: http://iieta.org/journals/rces

Performance Evaluation of Email Spam Text Classification Using Deep Neural Networks

Venkata RamiReddy Chirra[1*], Hoolda Daniel Maddiboyina[1], Yakobu Dasari[1], Ranganadhareddy Aluru[2]

[1] Department of Computer Science & Engineering, VFSTR, Guntur, India
[2] Department of Biotechnology, VFSTR Guntur, India

Corresponding Author Email: chvrr_cse@vignan.ac.in

https://doi.org/10.18280/rces.070403

Received: 3 November 2020
Accepted: 12 December 2020

Keywords:
RNN, LSTM, GRU, Bidirectional, NLP

ABSTRACT

Spam in email box is received because of advertising, collecting personal information, or to indulge malware through websites or scripts. Most often, spammers send junk mail with an intention of committing email fraud. Today spam mail accounts for 45% of all email and hence there is an ever-increasing need to build efficient spam filters to identify and block spam mail. However, notably today's spam filters in use are built using traditional approaches such as statistical and content-based techniques. These techniques don't improve their performance while handling huge data and they need a lot of domain expertise, human intervention and they neglect the relation between the words in context and consider the occurrence of the word. To address these limitations, we developed a spam filter using deep neural networks. In this work, various deep neural networks such as RNN, LSTM, GRU, Bidirectional RNN, Bidirectional LSTM, and Bidirectional GRU are used to a built spam filter. The experimentation was carried out on two datasets, one is a 20 newsgroup dataset, which contains multi-classes with 20,000 documents and the other is ENRON, a dataset contains 5,000 emails. The custom-designed models have performed well on both benchmark datasets and attained greater accuracy.

1. INTRODUCTION

Every second, around 2.4 million emails are sent across the world and email continues to remain as the communication medium for people throughout the world. Spam filter is much needed as it can shift out inapt or undesirable messages from unsure senders. The effort is to build an efficient filter for spam with high precision and an improved spam filter compared to the filters used twenty years ago. The outdated approaches like whitelisting, crude filters etc., that were put forward by ISP failed spectacularly as they were found to be inexact and highly redundant. Later, statistical methods like Naive Bayes, Support Vector Machine [1, 2], and Linear Regression became favored as they produced better results than whitelisting or keyword filters. Neural networks like Radial Basis Function (RBF) [3, 4] and deep Neural Network [5] are also used for text classification.

For example, Naive Bayes is a classification technique based on the probability between the input features which are strongly independent. This leads to low precision and poor performance with the multi-class text classification where the prediction becomes difficult if it is a multi-classifier. But better results were achieved with the statistical methods on spam filtering data with binary labelled classes. Indeed with these limitations, the statistical methods form a powerful baseline to the recent models.

Text classification has become a challenging task in Natural Language Programming (NLP) and various methods have been proposed. For the text classification, the text needs to be represented in vector form with real number values, which are assigned randomly. Embeddings in NLP are mapping words in a large collection of vocabulary to the randomly assigned real vector values. As the classification is important, the vectors which are randomly assigned with real number values shouldn't have any relation between them as it might cause cumbersome while mapping the words in the vocabulary.

The main drawback of TF-IDF is considering the document whole for classification. Word embeddings approach is used to overcome this, where in each word is taken as vector. The word embeddings are of two types i.e., word2vec and Glove. The binary classification of the email is processed through word2vec.The continuous bag of words (CBOW) model is a variant of word2vec embedding that finds the required word from the surrounding context of words. In this, grouping the text file into words is performed. It records the frequency of each word in each document and finally assign with an integer id. Every unique word in vocabulary will represent a feature and the context is represented by multiple words for a given target words. The multi classification of email documents is carried out through Glove embeddings. Glove embeddings are used to find the semantic relation between the words in the text classification.

2. RELATED WORK

Spam an unnecessary message, which is sent through a computerized medium globally. Not only through email, but also different types of social media are being affected by the fraudulent data that has been sent or received. The statistical methods like Naive Bayes, TF-IDF and SVM count the word occurrence in the document. Subramaniam et al. [6] proposed a Naive Bayes algorithm on collection of spam emails from google Gmail account dataset that have attained 96.8%

accuracy. Drucker et al. [7] proposed SVM algorithm on sample emails dataset that have attained 90-95% accuracy. Abdulhamid et al. [8] proposed a machine learning algorithm on UCI Machine learning repository dataset and achieved 94.2% accuracy. DeBarr and Wechsler [9] proposed a Random forest algorithm on custom collection dataset. This model yields 95.2% accuracy. Zhang [10] have proposed a CNN with noise reduction module on Grumble and Ling-spam dataset, which yields 98% accuracy. Lyubinets et al. [11] propose a RNN with embeddings, which produced 88.2% accuracy on Arch Linux bug tracker and Chromium bug tracker dataset. Eugene and Caswell [12] proposed a CNN with small filter size on ENRON dataset that have attained 84% accuracy. Du and Huang [13] proposed a LSTM with attention on NLPCC2014 and Reuters dataset. This model performed well with an accuracy of 81.9%. Banday and Jan [14] proposed Naive Bayes and SVM on Nepali SMS dataset. This model performed well with an accuracy of 92.74%. Shahi and Yadav [15] proposed Naive Bayes, K-Nearest Neighbor, SVM, classification Bayes Additive Regression Tree on Real life dataset. This model performed well with an accuracy of 96.6%. Table 1 summarize various algorithms for classification of spam mails on various datasets and their performance.

Naive Bayes is a traditional technique which is mostly used for text classification with strong independent assumptions. This technique calculates the entire probability of the document that belong to different classes and it assigns the document with highest probability.

The word count statistics based algorithms such as TF-IDF plays a very important role in the context of NLP. This algorithm represents each sentence as scores of vectors that are determined by term frequencies. The score is determined in two directions– one determines the frequency of that term inside the document and the other determines the presence of the term across several documents. The score is calculated by multiplication of these two. The resulting data of this algorithm can be used with any supervised classification method such as softmax classifier.

Stochastic Gradient Descent (SGD) is a discriminative iterative learning algorithm for optimizing the objective function. The SGD Classifier stops when the maximum iteration i.e., level of convergence has reached. Previously, SGD algorithms have been converged with back-propagation algorithms in multilayer neural networks.

3. PROBLEM STATEMENT

Spam is unsure text sent over the web, particularly to a more number of internet users, in the name of advertising, shopping, communicating and spreading malware over the networks. Spam email or junk mail are increasing rapidly over 2007 and present in almost 85% of all emails. Table 1 shows the accuracies of existing methods.

Nowadays spam mails percentage has been raised to 95%. According to Spam cop statistics: Average spam: 2.7 per second, Max spam: 4.7 per second, Total reported (last year): 85,734,997. Getting out of unsure text is often considerable as they might consume a lot of your inbox space and effort when you start clearing these unwanted mails. Malware and viruses can indulge into these emails and the company's confidential information might be stolen by spammers.

Table 1. Accuracy of various existing approaches

Author	Algorithm	Datasets	Accuracy (%)
Subramaniam et al. [6]	Naive Bayes algorithm	spam emails from gmail	96.8
Drucker et al. [7]	SVM	Sample emails	90-95
Abdulhamid et al. [8]	Machine Learning algorithms	UCI Machine learning repository	94.2
DeBarr and Wechsler [9]	Randomforest algorithm	Custom collection	95.2
Zhang [10]	CNN	Grumble and Ling-spam	98
Lyubinets et al. [11]	RNN with embeddings	Arch Linux bug tracker and chromium bug tracker	88.2
Du and Huang [13]	LSTM with attention	NLPCC2014 and Reuters	81.9
Banday and Jan [14]	Naive Bayes, K-NN, SVM.	Real life data set	96.6
Shahi and Yadav [15]	Naïve Bayes, SVM	Nepali SMS	92.74
Eugene and Caswell [12]	CNN with small filter size	ENRON dataset	84

In this situation, introducing spam filter in every organisation email servers is very essential. The text classification is done for binary classification (for ENRON dataset) and multi class classification (For 20 Newsgroup dataset) in the context of Natural Language Processing and is implemented by deep learning models that have achieved greater testing and validation accuracy than the traditional statistical models [16-19].

4. PROPOSED METHOD

Deep learning an emerging technology have proven their capabilities in the scope of NLP tasks. The obtained precision determines which model works better in the deep neural networks. The text classification of binary and multi class classification through deep learning models gained higher accuracy than statistical methods like Naive Bayes, SVM etc., Now, the proposed work is related to two aspects of research (i) Supervised methods for spam detection, and (ii) Spam detection through advancements in Deep learning techniques for NLP applications. Deep learning models have remarkable performance in the area of Natural Language Processing (NLP). The binary class and multi-class text classification of emails is carried out by word2vec and Glove (global vector for word representation). Word2vec assigns a unique integer id to each word after finding the frequency of each word in the whole document. Glove embeddings captures sub-linear relationships in the vector space perform better in the word analogy tasks. In this work, the multi-class text classification of 20,000 email documents were implemented by neural architectures and the binary classification of 6,000 emails of ENRON dataset were implemented through traditional methods and neural networks. RNN, LSTM, GRU,

Bidirectional RNN, Bidirectional LSTM, Bidirectional GRU neural networks. These perform well with the multi class text classification whereas statistical methods perform well for a binary classifier.

Text classification is one of the most important problems of Natural Language Programming (NLP) research and variety of methods has been proposed for it. For the classification of text, the text needs to be represented in numerical form which is mostly done by assigning random vectors with real number values. Embeddings in NLP are mapping words or phrases from the vocabulary to vectors of real numbers.

Tokenisation of the text in emails has to be done before forwarding it to neural network model as shown in Figure 1. The token can be a word or a number or punctuation. Each word is considered as token in a sentence and each sentence is a token in a paragraph. The tokenized text is used as the semantic identifiers in taxonomy. Sequence padding is then carried out on preprocessed tokenized text. If the sequence is larger than the maximum length given in argument then there is a problem in variable length sequence prediction, the sentences have to convert into same length. This can be achieved by adding dummy values or by truncating the sequence to desired length. One hot encoding on the padded sequences is vector representation with all the elements as 0 except one, which has 1 as its value. The position of the word in vocabulary is used to insert 1 at required index in the one hot vector. One hot vector is used to convert sentence into features that need to be fed into neural networks classifier model.

The multi classification of email documents is carried out through Glove embeddings. Global corpus statistics are directly retrieved by the model. Glove an unsupervised algorithm, achieves this with a co-occurrence matrix and by using matrix factorization. In this the words are forwarded and get them encoded into vectors. Glove embeddings are used to find the semantic relation between the words in the text classification. Figure 1 shows the spam classifier model.

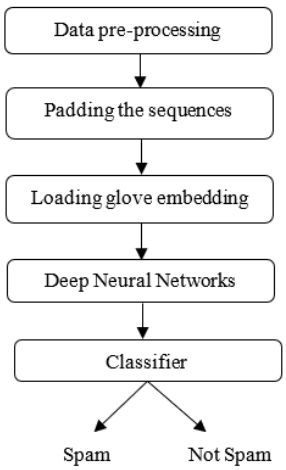

Figure 1. Spam classifier model

4.1 Deep Neural Networks

4.1.1 Recurrent Neural Network (RNN)

RNN has good learning ability in NLP tasks. It is characterized by good modelling of sequence of data and full utilization of sequence information. Simple RNN achieved 95.02% accuracy in 7 epochs.

4.1.2 Long Short Term Memory Neural Network (LSTM)

The traditional algorithms such as Naive Bayes, Logistic regression etc., can't memorize the past data. LSTM neural network, which is a variant of RNN has the ability to memorize the past data and can pass the previously obtained data in series of networks like architecture. LSTM has strong gradient over the time steps so that it can hold the long sequences. It comprises of a memory cell that holds the past data and 3 logic gates namely. Read gate reads the data from memory cell. Write gate writes the data into memory cell and Forget gate deletes the old data. These gates keep the LSTM networks away from the vanishing and exploding gradients. The main advantage of LSTM is it stores the information that is useful and deletes the data that is unnecessary. LSTM achieved 94.90% accuracy in 7 epochs.

4.1.3 Gated Recurrent Unit (GRU)

This is a variant of RNN which can overcome the problem of vanishing gradients and exploding gradients while handling the sequential data. The additional gates in GRU are reset and update gates and these gates add the functionality that it can store and filter using these gates. The update gate tells how much need to memorize the previous data. The reset gate tells how to merge the current input with the past value. When compared to LSTM, GRU'S are faster to train and need fewer data to generalize. GRU achieve 95.01%accuracy.

4.1.4 Bidirectional LSTM

This modification is done on normal LSTM networks to make it able to work better for NLP problems. Bidirectional runs the inputs in two ways, one from past to future and one from future to past. Bidirectional LSTM has no backward pass and it achieves about 94.65% accuracy.

4.1.5 Bidirectional GRU

Bidirectional GRU's are a type of bidirectional recurrent neural networks with only the reset and update gates. The prediction of the current state in this network is carried out by keeping track of the past and later time steps. It allows the use of information from both previous time steps and later time steps to make predictions about the current state. Bidirectional GRU attains greater accuracy of 95.01%.

4.2 Softmax classifier

The extracted features from the text called feature vectors are sent into softmax classifier for classification. It deals with multi class classification, softmax classifier works better than all other activation functions. Softmax is a kind of Multi Class Sigmoid, and softmax function is the sum of all softmax units is equals to 1 and the probability of each class lies in between 0 and 1. The softmax function is useful for converting an arbitrary vector of real numbers into a discrete probability distribution. The softmax function calculates the probability of all the n- classes in the classification. In classification task, for the given input features the class with high probability is predicted as the target class.

4.3 Dropout

The main drawback with neural networks is that when they need to work on the low volume of data the model overfits the data. So, the better thing to do is increasing the data and size

of the network. The model is optimized by just adding the dropout to the neural networks.

5. RESULT AND ANALYSIS

5.1 Datasets description

The 20 newsgroup is multi classified dataset and it reports 20,000 news group documents, divided into 20 different newsgroups. This dataset has become popular in text applications machine learning and deep learning techniques, such as text classification and text clustering. The Enron Email dataset consists of 5,000 emails. It is a binary classified dataset and is labelled as ham or spam. Table 2 shows the results of various deep neural networks formulate class classification and Table 3 shows the results for a binary classification.

5.2 Multi-class classification

Table 2 shows the accuracy of various models on 20 Newsgroup dataset. Comparison of validation and test accuracies of various deep networks is depicted in Figure 2.

Table 2. Accuracy all models on 20 Newsgroup dataset (%)

Model	Validation Accuracy	Testing Accuracy
Simple RNN	95.34	95.02
LSTM	95.89	94.9
GRU	96.9	95.01
Bidirectional RNN	95.4	95.02
Bidirectional LSTM	96.6	94.65
Bidirectional GRU	97	95.01

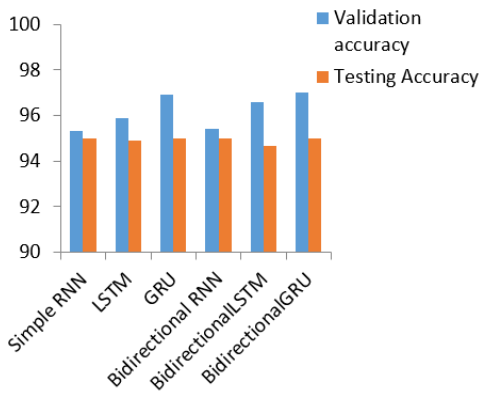

Figure 2. Comparison of validation and testing accuracies of various deep networks

5.3 Binary class classification

Table 3 shows the accuracy of various models on ENRON dataset. Comparison of validation and testing accuracies of various deep networks is depicted in Figure 3.

Table 3. Accuracy of various deep networks on ENRON dataset

Model	Testimg Accuracy	Validation Accuracy
CNN	98.5	99.2
LSTM	98.2	98.1
GRU	97.8	95.2

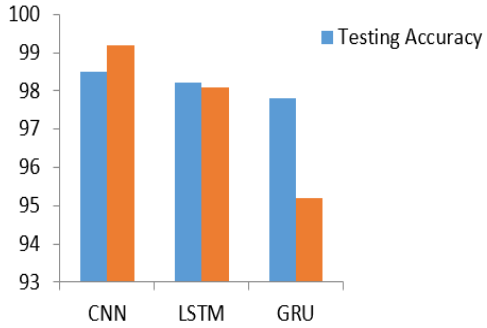

Figure 3. Comparison of testing accuracies of various deep networks

6. CONCLUSION

Detecting spam in social media is a search out problem that needs to keep an eye on junk and unwanted data. The text classification models were developed for building an efficient spam filter through several deep learning networks in the context of NLP. Deep learning networks and Glove embeddings for text classification performed well and obtained high testing accuracy even with large dataset. The greater accuracy of 95.02% was achieved for a multi-class classification and 98.5% was achieved for a binary classification.

REFERENCES

[1] Reddy, C.V.R., Reddy, U.S., Kishore, K.V.K. (2019). Facial emotion recognition using NLPCA and SVM. Traitement du Signal, 36(1): 13-22. https://doi.org/10.18280/ts.360102

[2] Ramireddy, C.V., Kishore, K.V.K. (2013). Facial expression classification using Kernel based PCA with fused DCT and GWT features. 2013 IEEE International Conference on Computational Intelligence and Computing Research, Enathi, pp. 1-6. https://doi.org/10.1109/ICCIC.2013.6724211

[3] VenkataRamiReddy, C., Kishore, K.V.K., Bhattacharya, D., Kim, T.H. (2014). Multi-feature fusion based facial expression classification using DLBP and DCT. International Journal of Software Engineering and Its Applications, 8(9): 55-68. https://doi.org/10.14257/ijseia.2014.8.9.05

[4] Reddy, C.V.R., Kolli, V.K.K., Reddy, U.S., Suneetha, M. (2016). Person identification system using feature level fusion of multi-biometrics. 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, pp. 1-6. https://doi.org/10.1109/ICCIC.2016.7919672

[5] Chirra, V.R.R., Uyyala, S.R., Kolli, V.K.K. (2019). Deep CNN: A machine learning approach for driver drowsiness detection based on eye state. Revue intelligence Artificial, 33(6): 461-466. https://doi.org/10.18280/ria.330609

[6] Subramaniam, T., Jalab, H.A., Taqa, A.Y. (2010). Overview of textual anti-spam filtering techniques. International Journal of Physical Sciences, 5(12): 1869-1882. https://doi.org/10.5897/IJPS.9000424

[7] Drucker, H., Wu, D., Vapnik, V.N. (1999). Support vector machines for spam categorization. IEEE

Transactions on Neural Networks, 10(5): 1048-1054. https://doi.org/10.1109/72.788645
[8] Abdulhamid, S.M., Shuaib, M., Osho, O., Ismaila, I., Alhassan, J.K. (2018). Comparative analysis of classification algorithms for email spam detection. International Journal of Computer Network and Information Security (IJCNIS), 10(1): 60-67. https://doi.org/10.5815/ijcnis.2018.01.07
[9] Debarr, D., Wechsler, H. (2009). Spam detection using clustering, random forests, and active learning. Sixth Conference on Email and Anti-Spam. Mountain View, California.
[10] Zhang, W. (2018). Spam filter through deep learning and information retrieval. Dissertation, Johns Hopkins University.
[11] Lyubinets, V., Boiko, T., Nicholas, D. (2018). Automated labeling of bugs and tickets using attention-based mechanisms in recurrent neural networks. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, pp. 271-275. https://doi.org/10.1109/DSMP.2018.8478511
[12] Eugene, L., Caswell, I. (2015). Making a manageable email experience with deep learning. 1-8.
[13] Du, C., Huang, L. (2018). Text classification research with attention-based recurrent neural networks. International Journal of Computers Communications & Control, 13(1): 50-61. https://doi.org/10.15837/ijccc.2018.1.3142
[14] Banday, M.T., Jan, T.R. (2009). Effectiveness and limitations of statistical spam filters. arXiv preprint arXiv: 0910.2540.
[15] Shahi, T.B., Yadav, A. (2013). Mobile SMS spam filtering for Nepali text using naïve Bayesian and support vector machine. International Journal of Intelligence Science, 4(1): 24-28. https://doi.org/10.4236/ijis.2014.41004
[16] Banothu, B., Murthy, T.S., Reddy, C.V.R., Yakobu, D. (2020). High-order total bounded variation approach for gaussian noise and blur removal. International Journal of Advanced Science and Technology, 29(3): 10152-10161.
[17] Yakobu, D., Reddy, C.V.R., Sistla, V.K. (2019). A novel energy efficient scheduling for VM consolidation and migration in cloud data centers. Ingénierie des Systèmes d'Information, 24(5): 539-546. https://doi.org/10.18280/isi.240512
[18] Bulla, S., Reddy, C.V.R., Padmavathi, P., Padmasri, T. (2020). Analytical evaluation of resource estimation in web application services. Ingénierie des Systèmes d'Information, 25(5): 683-690. https://doi.org/10.18280/isi.250516
[19] Rekha, S.N. (2014). A review on different spam detection approaches. International Journal of Engineering Trends and Technology, 11(6): 315-318. https://doi.org/10.14445/22315381/IJETT-V11P260