Email Spam Detection Using Multilayer Perceptron Algorithm in Deep Learning Model

Senthil Murugan Tamilarasan, Muthyala Hithasri, and Kamakshi Pille

Abstract Email spam detection is a filtering process which identifies either it is spam or not. It also removes the unsolicited data present in the user's email inbox. Certain type of spam mails contains malware which misuse the users' data. Hence, we need to identify spam mails and take necessary actions. Many machine learning algorithms have proposed for differentiate spam mails from normal mails. Tokenization of emails between length and frequency is one of the techniques. It helps to split the raw emails into tokens known as small words. After tokenization, tokenized count has taken into consideration for process the emails. Based on that spam emails to be identified which are present in the dataset of spam and ham emails. To extracting these features, term frequency—inverse document frequency (TF-IDF) has used to train the model. In this method, multilayer perceptron deep learning algorithm is applied to compute the model. It has two layers. When input is given to the perceptron, the input is multiplied by the hidden layers, and it holds the activation function such as sigmoid activation with regularization function. For the better optimization, the model uses the Adam optimizer with gradient descent for fastest optimization. The network learns the model. The learning rate is set to true. While computing the model, it goes in the forward direction to train the model and comeback again (backpropagation). This process will be repeated. Going to the forward direction and comes back, then again, maintaining forward approach is called one epoch. The epoch rate has computed in the model. In the comparison between multilayer perceptron algorithm and machine learning algorithms such as support vector machine (SVM), random forest, and XGBoost, the deep learning algorithm produces 99% of accuracy on precision, recall, and F-measure and holds less computation time. Hence, the results prove that deep learning algorithm performs better than machine learning algorithms.

Keywords Email spam detection · Deep learning model · Multilayer perceptron

S. M. Tamilarasan · M. Hithasri (✉) · K. Pille
Kakatiya Institute of Technology and Science, Warangal, Telangana, India
e-mail: hithasri.muthyala@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

A. Joshi et al. (eds.), *Information and Communication Technology for Competitive Strategies (ICTCS 2021)*, Lecture Notes in Networks and Systems 400,
https://doi.org/10.1007/978-981-19-0095-2_55

1 Introduction

Spam mails may be sending to user's inbox directly. Unfortunately, user may open the mail or click some links which is in that mail. It may cause for stoles the users' personal data by frauds. In these days, apart from email, certain type of spam is increasing through text message services. Such type of spamming occurs in various categories in many ways. In this paper, we are going to prepare spam filter and identifying them either spam or not spam. Previously, various types of machine learning algorithms were used for detection of spam mails. But, we are going to identify spam through deep learning model with the techniques of multilayer perceptron. We expect to achieve 99% accuracy in performance, f1-score, and recall. The detection takes place by many of the processing techniques like importing of libraries and preparing of Enron dataset which easily distinguishes between ham and spam, and the filter consists of many layers, and few types of dictionaries are used.

Khan et al. [1] proposed TF-IDF model is created for extracting the features from tokenizing the input elements present in the data. Keras software tool acts as front end, and TensorFlow act as backend. Layer1 (l1) and layer2 (l2) consist of neurons with linear ReLU, with sigmoid activation function. Sigmoid is an activation function present in the deep neural networks. Regularization is a technique reduces the complexity in neural networks. Here, regularization function is used to solve more complex problems. Dropout is taken with probability rate is used to prevent the problem of over fitting in l1 and l2. Loss function is implemented on binary cross-entropy and for fastest optimization in gradient descent estimation.

Gradient descent is an optimization technique. Adam optimizer is used to better optimization and data shuffling which helps to improve the performance of the model. This model has chosen batch size and learning rate. If the learning rate increases when compute the model, it takes more time which is present in the deep learning. Features are extracted automatically in the deep learning model. Deep learning takes more time when compared to the machine learning. But, the problem with the machine learning is those features that are not trained automatically. We have to train the model. We should give some type of inputs. Machine learning algorithms are hands-crafted one, while deep learning model takes automatically trained the model.

2 Literature Survey

In machine learning techniques, the following methods are used to find the spam mails. Such as support vector machine, random forest, decision tree, and Naive Bayes. Also, those techniques have been compared with bio-inspired metaheuristic algorithms. Some genetic algorithm and particle swarm optimization algorithm are also used to detect spam mails. These algorithms are implemented using seven different types of datasets and different tools such as WEKA, scikit-learn, Keras, TensorFlow. Srinivasan et al. [2] used Python platforms on Jupiter Notebook which is an

open-source tool and also used similar to Spyder ide. In Spyder, the programmer can capable to include multiple panels such as console where the output can be visible and variable explorer which assignment of variables. Online platforms such as Google Collaboratory, Kaggle desktop top-based platforms are also used.

The Enron-Spam dataset contains the following six datasets. SpamAssassin dataset converts folder into text files. Gibson et al. [3] introduced a Ling-Spam dataset contains ten parts where dataset is not preprocessed. PUA is a type of numerical dataset that has different types of mails. Dataset has been splitted into different ratios like 60:40, 70:30, 75:25 and 80:20 for training and testing respectively. Scikit-learn is the library they had used to measure the performance. Recall, precision, f1-score, accuracy had been calculated using confusion matrix, on every dataset. Some of the tuning parameters has been applied by Gangavarapu et al. [4] for detection of emails like SGD, MNB, MLP, DT, and RF. A particle swarm optimization (PSO) which is bio-inspired algorithms considers the best evaluation position and global position for usage of machine learning models. It is used in PSO for the purpose of feature selection. Number of particles also used as input for this model and it is considered from library.

Genetic algorithm iterates through a fitness function between two individuals and produces offspring. They have been tested the models with the 50,000 emails. Awad and Foqaha [5] tested on the numerical corpus, restrictions occurred in the extracting the features. While they implemented on alphabetical corpus, it had given the best results in extracting the features and predicting outcome. WEKA acts as a black box that ran the datasets on 14 different classification algorithms. It produces best results in the top algorithms like multinomial Naive Bayes, support vector machine, random forest, and decision tree. All these algorithms have been tested and experimented with scikit-learn library and its related modules by Hussain et al. [6].

Genetic algorithm works better on the text-based datasets as well as numerical datasets than PSO. PSO worked efficiently with the multinomial Naive Bayes and stochastic gradient descent. Whereas, genetic algorithms worked well for decision tree and random forest. On overall results, Naive Bayes algorithm works better for detection of spam detection of emails, and multinomial Naive Bayes performed highest accuracy of 98% with genetic algorithm optimization for 80:20 split of train and test on spam assassin dataset.

3 Machine Learning Models

Three models like random forest, support vector machine, and XGBoost extracting the TF-IDF feature have been tested for compare the results. Many researchers have done on the machine learning models like Naive Bayes support vector machine, decision tree, XGBoost, random forest for spam detection.

3.1 Support Vector Machine (SVM)

In this method, train the classifier to take learning features and learn how they relate to the training of species and iterates. It has done over the linear data and sample data to optimize linearly in a hyper-plane stochastic gradient descent.

3.2 Random Forest

Random forest algorithm is used for both classification and regression model implement on scikit-learn library. It uses on termination criteria means; it goes with deep depth for computation process. It takes more time. Evaluation goes on random forest model and selects the higher number of classes as prediction. Random forest is an algorithm depends on the decision tree. If tree is complex, then go for the random forest algorithm where decision tree goes in an top-down approach manner.

3.3 XGBoost

XGBoost is known as extreme gradient boosting; it implements on the gradient-boosted decision trees designed for speed and performance. It can also be used for classification and regression model in training the model. XGBoost is taken as XGBoost and fit model.

4 Different Steps in Deep Learning

This section describes the step-by-step processing of deep learning techniques.

4.1 Preprocessing the Data

Preprocessing results in creating an Enron dataset and generate pandas' data frame. Some analyses are applied on the data like tokenizing. First step of model is importing the libraries and preparing an Enron dataset into the framework, load, and extract the Enron data spam in panda's framework. After loading the Enron data, data frame is applied for tokenized text, tokenized count, and language.

4.2 Preparation of Training and Test Data

Split the dataset for training and validation. We are going to take approximately 10,000 emails for testing, and remaining emails take place for build the model. The data are going to shuffle in the data frame.

4.3 Feature Generation

In the process of feature generation, unsupervised learning algorithm is converted to the supervised learning algorithm using the labeled Enron-Spam dataset. TF-IDF models are used for counting number of spam words, stop words, repeating words held in the dataset for training the models. Importing Keras library in TensorFlow acts as back end and prepare the function using TF-IDF feature models for generation of input. This model has taken totally 34 s for building the model.

4.4 Split and Validation of Data

This step is useful in order to prevent the model from overfitting. Here, split the data for training and testing. Collecting a sample of data remained back from the training model results in validation phase and estimating it in tuning the models in hyper-parameters present in the data. Hence, the entire dataset has been splitted into 85:15 ratio for training and testing respectively. We are importing test, train, and split for validation, and model selection is used from the Sklearn library.

4.5 Model Building

In this model building, it takes two input layers as layer one consists of linear ReLU, and layer two consists of linear ReLU with sigmoid function which is an activation function used to predict the probability of the model. This sigmoid activation function is mostly used in the multilayer neural networks. It predicts in shape ‘s’ which is a nonlinear activation function. Layer one consists of 512 neurons, and layer two consists of 256 neurons with sigmoid activation function to avoid the problem of overfitting. A regularization function is used dropout rate and also added for it with 0.5 probabilities in l1 and l2. Loss function is used for back propagating the model in a sequence manner. Loss function is applied on the binary cross-entropy to get optimization for faster estimation.

Adam optimizer is using gradient descent estimation to perform accurately on parameters and to minimize the cost function. Data shuffling meant that shuffle the

data within an attribute maintained at true and batch size is taken as 64. The network has initialized with 0.001 learning rate, and check points were kept spam detection of spam takes place.

5 Results

Deep learning model with multilayer perceptron does very well on the test data. The results from other models are close but more. We tried this approach over multiple language emails, and deep learning model is very consistent with the performance. XGBoost also does very well.

In confusion matrix, true-positive rate means when the model is predicted true and output also predicted as true. True-negative rate means when the model is predicted positive and output is predicted as negative. False-positive rate means when the model is false and also output predicted as negative and false positive rate means when the model is false and output is predicted as negative (Table 1).

Table 2 shows the classified outputs in the following rates are true positive and false negative because they are results in the same as present in the input. These rates are used in the data-related topics for finding the accuracy how much actively our preferred type of model is performing related to our approach on the confusion matrix; these types of values are predicted; confusion matrix is divided into four blocks in which each size has a rate such as true positive, true negative, false positive, and false negative using this confusion matrix; only, the predicted output is generated perfectly by using the model extraction. F1-score is also called as f-measure where f is a randomly chosen. Precision finds the accurateness of values the relevant instances which are retrieved among the instances, and recall is also called as the sensitivity

Table 1 Confusion matrix

Model	False negative	True positive	True negative	False positive
Random forest	732	5187	3743	338
SVM	540	5379	4026	55
XGBoost	479	5440	3398	683
Deep learning	46	5873	4037	44

Table 2 Comparison results based on precision, recall, F1-score, and samples

Model	Precision	Recall	F1-score	Total samples
Random forest	0.887624	0.896754	0.890721	10,000
SVM	0.935807	0.947646	0.939390	10,000
XGBoost	0.882452	0.875857	0.878744	10,000
Deep learning	0.990649	0.990723	0.990686	10,000

used find the relevant instances that are retrieved actual instances.

$$\text{Precision} = (\text{True positive})/(\text{True positive} + \text{false positive}) \quad (1)$$

$$\text{Recall} = (\text{True positive})/(\text{True positive} + \text{false negative}) \quad (2)$$

$$F1\ \text{score} = 2 * ((\text{Precision} * \text{recall})/(\text{Precision} + \text{recall})) \quad (3)$$

6 Conclusion

We concluded that deep learning model with the multilayer perceptron algorithm produces the accuracy of 99% in precision, recall, F1-score. The final results are compared with machine learning algorithms such as random forest, SVM, and XGBoost, and it is shown in Table 2. We considered only 1000 samples from Enron dataset for the testing and generated confusion matrix which shown in Table 1. Finally, deep learning algorithm produced more good results compared with machine learning algorithms. It may be vary when using different datasets and total number of data at the time of tuning. In future, the deep learning researchers may apply bio inspired based algorithm for improving the results in this model.

References

1. Khan WZ, Khan MK, Muhaya FT, Aalsalem MY, Chao HC (2015) A comprehensive study of email spam botnet detection. IEEE Commun Surv Tutorials 17(4)
2. Srinivasan S, Ravi V, Alazab M, Ketha S, Al-Zoubi AM, Kotti Padannayil S (2021) Spam emails detection based on distributed word embedding with deep learning. In: Maleh Y, Shojafar M, Alazab M, Baddi Y (eds) Machine intelligence and big data analytics for cybersecurity applications. Studies in computational intelligence, vol 919. Springer, Cham
3. Gibson S, Issac B, Zhang L, Jacob SM (2020) Detecting spam email with machine learning optimized with bio-inspired metaheuristic algorithms. IEEE Access 8:187914–187932. https://doi.org/10.1109/ACCESS.2020.3030751
4. Gangavarapu T, Jaidhar CD, Chanduka B (2020) Applicability of machine learning in spam and phishing email filtering: review and approaches. Artif Intell Rev 53:5019–5081
5. Awad M, Foqaha M (2016) Email spam classification using hybrid approach of RBF neural network and particle swarm optimization. Int J Netw Secur Its Appl 8(4):17–28
6. Hussain N, Mirza HT, Hussain I, Iqbal F, Memon I (2020) Spam review detection using linguistic and spammer behavioural methods vol 8. IEEE Access 2020