Comparison of Machine Learning and Deep Learning Methods for Spam Email Classification

Ashish Negi[1] Ginni Sharma[2] Jyoti Vishwakarma[3] Neha Pandey[4] Prof. Vikas S. Kadam[5]
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]Sinhgad Institute of Technology, Lonavala, India

Abstract— Email is one of the most popular communication medium in today's technology and is widely used by many people. How-ever, spam emails as become a significant problem in today's society causing tremendous losses at individual as well as organizational level. Various email providers today are using different techniques to classify emails as spam or ham how-ever, spammers continue to develop new techniques to outsmart them. In this research paper, our agenda is to find out the most efficient algorithm for text-based spam emails that contain both the subject and contents of the email. Here, we also use different Pre-processing techniques on our email text before executing the algorithms so that they give greater accuracy. We also provide performance comparison between the various machine learning and deep learning techniques and measure them mainly on three factors i.e. Spam recall, Spam precision and Accuracy.

Keywords: Spam, Spam emails, Spam email detection, Deep learning, CNN, LSTM, BiLSTM, Hybrid (BiLSTM +CNN)

I. INTRODUCTION

Emailing is one of the most convenient and ubiquitous forms of communication for professional and personal use. It is also the fastest way to send complex information from one user to another, including not just text but also attachments such as images, videos, documents and URLs. This form of communication helps us save a lot of time and cost by eliminating overheads.[1] In fact, it one of the most popular, cheapest and fastest means of communication in recent times. Since its inception, email has taken global operations to a new level of accelerated economic growth.

Despite the large amount benefits of emails unfortunately it has a very large presence of unsolicited and fraudulent emails. [4] The impact of spam email is less serious on an individual level but more serious from an organizational level. Sending spam not only affects email recipients of productivity and privacy, it also threatens their trust and reputation. An individual email user suffers a loss of 1.4 percent of their productive time in managing spam emails. Also, spam email costs businesses monetary losses and decrease workforce productivity. Spam emails also contain viruses or URL link in the messages that might point to the website containing viruses which will damage the computer and as a result cause a problem to the company which may lead to the loss of trust of internet users using email technologies. [7]

An efficient Spam detection system is needed in today's world to protect email users from several negative usages to which it has been subjected recently. Unfortunately, due to the adaptive nature of unsolicited emails through the use of mailing tools and the false positive rate of anti-spam technology, the effectiveness of the spam detecting tools have often been limited and sometimes rendered ineffective or compromised, hence the need for better spam detection tools to achieve better spam detection accuracy.[4]

Although there is a continuous effort by researchers and practitioners to develop spam detection systems with high accuracy, Email users still receive vast amounts of spam messages every day. With help of spam botnets, malware and spam campaigns spammers can send thousands of messages at low or no cost at all.

In the entire technology development process, there are mainly three types of technical methods are used for spam detection including blacklisting, Classification algorithm based on machine learning, and deep learning. The current spam detection methods are based on blacklisting mechanism mainly rely on people for identifying and reporting phishing links requires a large amount of manpower and time. However, applying a spam detection model based on a machine learning classification algorithm requires feature engineering to manually find the appropriate features that are not likely to be affected by the migration of the application. The current spam detection method based on deep learning is limited to using word embedding for the content of the email. Legitimate message-sender has also sought to enrich the email nowadays by adding multimedia content like images which are used to hide fraudulent messages. Understanding the increasing trend of multimedia enriched email messages, it is the need of time to use such information to defeat the spammer's intention.

II. LITERATURE REVIEW

In 2017 Aakash Atul Alurkar, Sourabh Bharat Ranade, Parikshit N. Mahalle, Shreeya Vijay Joshi, Siddhesh Sanjay Ranade, Piyush A. Sonewa, and Arvind V. Deshpande [1] proposed the classification of emails as spam and ham using a machine learning approach which facilitated the algorithm to recognize the necessary features more accurately, rather than specifying them manually. The main idea of this model was to classify the incoming emails for a user using various parameters which are typically used by spammers. This paper presents a model which uses machine learning using parameters such as - To field, From field, Message-ID, Cc/Bcc field, etc. in the email header. The research paper mainly focuses on the email body with commonly used keywords and punctuations.

In 2013, Shukor Bin Abd Razak [2] had highlighted several features contained in the email header which can be selected to identify and classify spam messages efficiently. This paper considers the feature contents in Yahoo mail, Gmail, and Hotmail so that a generic spam messages detection mechanism could be proposed for all major email providers. This paper has presented an analysis of several features in the email header that could efficiently detect spam messages. Last but not least, this paper proposed potential features which are applied by the spammer to create spam

messages to be used in the spam detection system to provide efficient filtering of spam messages.

In 2015, Sunil B. Rathod, Tareek M. Pattewar [3] proposed a technique for email spam detection using only the content of the email. The model used the Bayesian approach for classifying Spam and legitimate mails using supervised learning across features extracted and by applying the Bayesian classifier. This model has experimentally demonstrated it can identify spam emails with an accuracy of 96.46% on real world Gmail data sets.

In 2019, Olatunji [4] has presented the Support Vector Machines-based technique for spam detection when presenting attention for successful search to the optimal parameters for obtaining the best performance. The tests were performed on both training and testing datasets. The outcomes have shown that the suggested approach was superior to existing approaches.

In 2016, Gauri Jain, Manisha, Basant Agarwal [5] have recommended a new predictive approach on the basis of deep learning techniques such as RNN and CNN for Spam Detection in Social Media. They have identified deep learning techniques which are more effective in text classification problems like spam detection because these models directly work on the raw data by learning high level features on their own. This is useful in the case of social media as no language structure is followed there and the people tend to use slangs and short forms.

In 2019, Yong Fang, Cheng Zhang, Cheng Huang, Liang Liu, AND Yue Yang [6] had used a deep learning model named THEMIS to detect phishing emails. The model used an improved RCNN to model the email header and the email body at both the character level and the word level which results in less noise in the model. The model used the attention mechanism in the header and the body, making the model pay more attention to the more valuable information between them. This model also used unbalanced dataset closer to the real-world situation to conduct experiments and evaluate the model. The THEMIS model obtains a promising result. The experimental results showed that the overall accuracy of the THEMIS model is 99.848%.and the false-positive rate (FPR) is 0.043%. The High accuracy and low FPR ensure it can identify legitimate emails and can identify phishing emails with high probability. This is one of the most promising results and is superior to the existing spam detection methods which verifies the effectiveness of THEMIS in detecting phishing emails.

III. METHODOLOGY

This paper proposes the classification of emails as spam and ham using a machine learning approach which facilitates the algorithm to recognize the necessary features more accurately, rather than specifying them manually. The main idea is to classify the incoming emails for a user using various parameters which are typically used by spammers. Its main objective is to group important emails and block spam ones. In this paper, we compare the effectiveness of email spam detection various algorithms of machine learning and deep learning model using the parameters such as email header and email body.

We have divided our model into four phases. The first phase comprises Data Acquisition and Data Preprocessing where the labelled datasets are utilized and preprocessed. The labelled and is preprocessed by performing some Natural Language Processes (NLP) such as stop word and punctuation removal, converting into lowercase English letters, and lemmatization.

The second phase deals with the feature selection process that includes TFIDF, Count Vectorizer, and Word Embeddings (Word2Vec) techniques. For traditional machine learning techniques, we have applied both TF-IDF and Count vectorizer techniques and for deep learning methods we have used Word Embeddings (Word2Vec) techniques for CNN, LSTM, BiLSTM, and Hybrid model (BiLSTM+CNN) to represent texts as numerical values.

The third phase is the spam detection phase where both machine learning and deep learning classifiers are used to distinguish email as spam and ham. Support Vector Machine (SVM), Decision Trees, and Naive Bayes (NB) classifiers are applied to detect spam reviews on behalf of traditional machine learning techniques. Convolutional Neural Network (CNN), LSTM (which is a variant of RNN), BiLSTM, and a combination of BiLSTM and CNN models are the deep learning techniques we have used for spam detection.

In the fourth phase, we have finally compared the performance of each classifier for both traditional machine learning and deep learning.

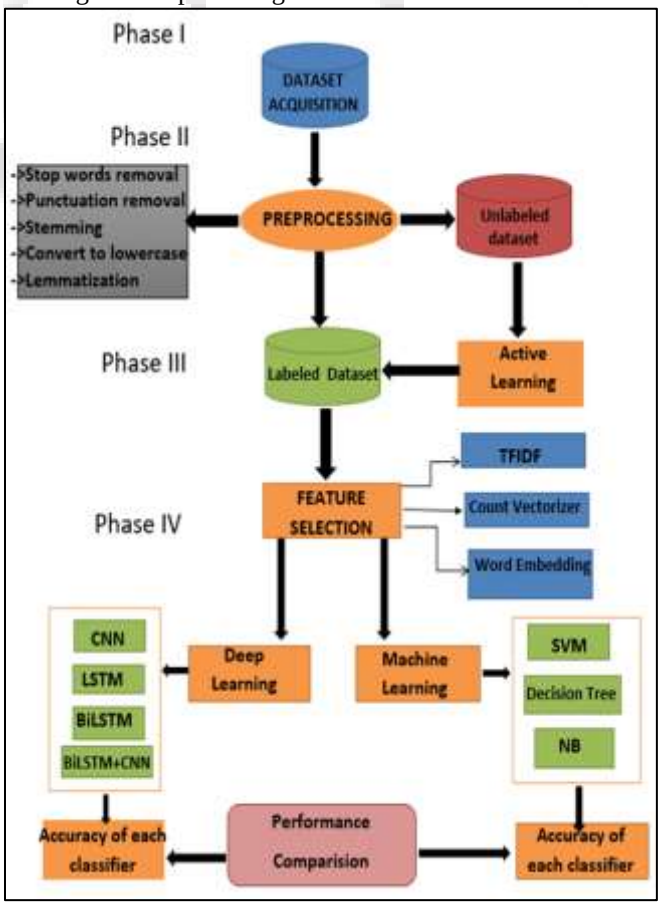


Fig. 1: Proposed Model for Email Spam Detection

A. The Algorithm for our proposed model can be inscribed in Algorithm 1.

Algorithm 1

```
1) loadData();
2) preprocessData();
3) shuffleData();
4) splitData();
5) wordEmbedding();
6) loadModel();
7) for each epoch in epochNumber do
8) for each batch in batchSize do
9) logit = model(feature);
10) loss = crossEntropy(logit, target);
11) loss.backward();
12) Evaluation(trainData, model, bestTrainAccuracy);
13) Evaluation(testData, model, bestTestAccuracy);
14) end
15) end
16) Function Evaluation(data, model, best):
17) correct = 0;
18) for each batch in batch Size do
19) logit = model(feature);
20) loss = crossEntropy(logit, target);
21) correct += (predict.data == target.data ).sum();
22) end
23) accuracy = (correct/totalData)*100;
24) best = max(best, accuracy);
25) return
```

1) Phase I: Preprocessing

In our model Data Acquisition and Data Preprocessing are discussed in Phase I. A detailed explanation of each of them is discussed below.

1) Data Acquisition: There are a handful of datasets that contain ham emails and also spam emails. Mostly labelled datasets have been used and they have been obtained from the following sources.
   Dataset table:

Dataset	Number of emails
Enron	33690
kaggle	2857
Lingspam	2591
Spam assassin	5169
Spam detection	2872
Trec07	4135

Total numbers of emails: 51314

2) Data Preprocessing: The labelled datasets are then pre-processed. Preprocessing is done by performing using Natural Language Processes (NLP) which includes the following steps 1. Tokenization 2. Lowercasing English letters 3. Punctuation removal 4. Stop words removal 5. Lemmatization.

2) Phase II: Feature Selection

The second phase of our proposed model is feature selection. Three feature selection techniques are used in our proposed model such as count vectorizer, TF-IDF and word embeddings. Feature selection for traditional machine learning classifiers is performed using TF-IDF values and count vectorizer techniques. Word2Vec word embeddings procedure is used for feature selection in the case of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), Bidirectional LSTM, and BiLSTM with Convolutional Neural Network (CNN) model.

1) Count Vectorizer: It converts the given text into a vector on the idea of frequency (count) of every word that happens within the entire text.
2) TF-IDF: Term frequency and inverse document frequency is a combination of two statistics.
   1) Term Frequency (TF) is the frequency of the word in the current document.
   2) Inverse Document Frequency (IDF) is the score of the words among all the documents.

   These scores highlight the words that are unique that is the words that represent important information in a respective document.
3) Word Embedding: Word Embedding is a technique where each word is basically represented as a vector of some dimension instead of a single number, here the semantic information and relation between different words is preserved by giving closeness in the vector space to words having the similar meaning of words.

3) Phase III: Spam Review Detection

To detect spam reviews, we have used three traditional machine learning classifiers such as Naive Bayes (NB), Decision trees and Support Vector Machine (SVM). Also, four deep learning techniques Convolutional Neural Network (CNN),), LSTM (which is a variant of RNN) ,BiLSTM and a combination of LSTM and CNN models are used for classification in order to compare classification performance with Machine Learning classifiers.

1) Convolutional Neural Network (CNN): In our work, for the spam assassin dataset, a convolutional neural network is used and significant results are achieved that outstand the performance of traditional classifiers. The whole process for in the simplest case, the whole process of CNN can be described for a single review. At first, the review text is pre-processed by performing NLP techniques and then the input texts are represented as a matrix. Each row of the matrix is a vector that represents a word and these vectors are known as word embeddings. The embedding layer will map the different words of the vocabulary. In our case, vocabulary size is the number of distinct words in our comments, and 100 means that we are going to represent each word with the vector of size 100. Then we execute a convolution by sliding a filter or kernel over the input that produces a feature map. At every location, matrix multiplication is performed and the result is summed up onto the feature map. In our work, we have used a kernel or filter regions of sizes 4 and each region has 100 filters. For introducing non-linearity, the output of the convolution is passed through the ReLU activation function. Next, to continuously reduce the dimensionality of the number of parameters and computation in the network we have used Max Pooling as a pooling layer that reduces the training time and controls over fitting. The whole process is repeated a number of times and finally, the result of the previous operations is passed to the Fully Connected Layer is to use these features for classifying the input data into two classes - spam or ham based on the training dataset.

2) Long short term memory(LSTM) /Bidirectional LSTM: Recurrent neural networks, and specifically LSTM and

BiLSTM, are known for their ability to keep the chronological order between data, which is very important when analyzing long-text opinions. LSTM/BiLSTM is different from CNN, thus it does'nt require a model to create input vectors.

The LSTM/BiLSTM models were configured as follows:

- Vocabulary size of 5000.
- LSTM/BiLSTM layer with 100 memory units.
- Since it's a classification problem Dense output layer is used with a single neuron.
- Sigmoid function is used as an Activation function.
- The loss function used is "binary-crossentropy" as it's a binary classification problem.
- Then "Adam" optimizer in used.

3) Hybrid model (CNN + BiLSTM): This model combines two neural networks, namely CNN and BiLSTM. We combined these models to test the adaptation of CNN with BiLSTM. The speciality of this model is that it allows extracting the utmost amount of data from documents using CNN convolution layers. This output becomes the BiLSTM input, which allows keeping the chronological order between the info present in both directions.
   - The combination of CNN and BiLSTM models requires a particular design since each model has a specific architecture and its strengths:
   - The ability of CNN is to extract as many features as possible from the text.
   - LSTM/BiLSTM keeps the chronological order between words in a document, thus it can ignore unnecessary words using the delete gate.

   The purpose of combining these two models is to build a model that takes advantage of the strengths of CNN and BiLSTM, in order that it captures the features extracted using CNN, and uses them as an biLSTM input. Therefore, we develop a model that meets this objective, such the vectors inbuilt in the word embedding part are used as convolutional neural network input. For each filter, a layer of max-pooling is applied to update and reduce the size of the data.

   Then, the results of all max-pooling layers are concatenated to build the BiLSTM input, which applies a BiLSTM layer to filter the information, using its three gates. The output of this step is that the input of the fully connected layer, which links each bit of input information with a bit of output information. Finally, we apply the sigmoid function as an activation function to assign classes to articles to produce the desired output.
4) Traditional Machine Learning: Decision trees, Naive Bayes (NB) and Support Vector Machine (SVM) classifiers are used as traditional machine learning techniques for spam detection for all the datasets. We have tried Count vectorizer and TF-IDF and listed the best accuracies to compare the performance of traditional classifiers and deep learning classifiers for the Dataset.

4) Phase IV: Performance Evaluation

In order to check the performance of the above mentioned six methods, we used the formost popular evaluation methods employed by the spam filtering researchers. Spam Precision (SP), Spam Recall (SR), Accuracy (A). Spam Precision (SP).

Spam Precision (SP) is the number of all spam emails that are correctly classified as spam to a number of all documents classified as spam. This shows that the noise that filters present to the user (i.e. no. of messages classified as spam will actually be spam).

$$SP = \frac{\text{\# Spam correctly Classified}}{\text{Total \# of messages classified as spam}} = \frac{N_{spam\to spam}}{N_{spam\to spam} + N_{ham\to spam}}$$

$$SP = \frac{\text{\# of Spam Correctly Classified}}{\text{Total \# of messages classifies as spam}} = \frac{N_{spam\to spam}}{N_{spam\to spam} + N_{ham\to spam}} \quad (8)$$

Spam Recall (SR) is the percentage of all spam emails that are correctly classified as spam.

$$SR = \frac{\text{\# of Spam Correctly Classified}}{\text{Total \# of messages}} = \frac{N_{spam\to spam}}{N_{spam\to spam} + N_{spam\to ham}}$$

$$SP = \frac{\text{\# Spam correctly Classify}}{\text{Total \# of messages}} = \frac{N_{spam\to spam}}{N_{spam\to spam} + N_{spam\to ham}}$$

Accuracy (A) is the percentage of all emails that are correctly categorized.

$$A = \frac{\text{\# of e-mails correctly categorized}}{\text{Total \# of emails}} = \frac{N_{ham\to ham} + N_{spam\to spam}}{N_{ham} + N_{spam}}$$

Where $N_{ham\to ham}$ and $N_{spam\to spam}$ are the number of messages that have been correctly classified to the legitimate email and Spam email respectively; $N_{ham\to spam}$ and $N_{spam\to ham}$ are the number of legitimate and spam messages that have been misclassified; N_{ham} and N_{spam} are the total number of legitimate and spam messages to be classified.[1]

IV. Results Analysis and Comparison

In this section, we summarize the performance result of the three machine learning methods and four deep learning methods in term of spam recall, precision and accuracy. Table 1 summarizes the results of the machine learning and deep learning algorithms in terms of Spam Precision on the different datasets whereas Table 2 summarizes the results of the machine learning and deep learning algorithms in terms of Spam Recall on the different datasets. Finally, Table 3 summarize the results of various machine learning and deep learning algorithms in terms of Spam Precision, Spam recall and Spam Accuracy on the different datasets. From the Table 3 we conclude that the Hybrid model (BiLSTM + CNN) method is most consistent in terms of performance on all three parameters i.e. of Spam Precision, Spam recall and Spam Accuracy closely followed by Bi LSTM and CNN. We can also conclude that LSTM has the least spam accuracy, precision and recall.

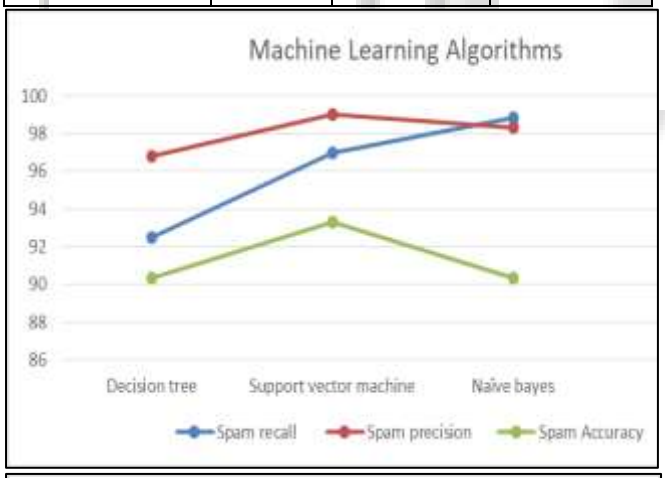
Spam precision table:

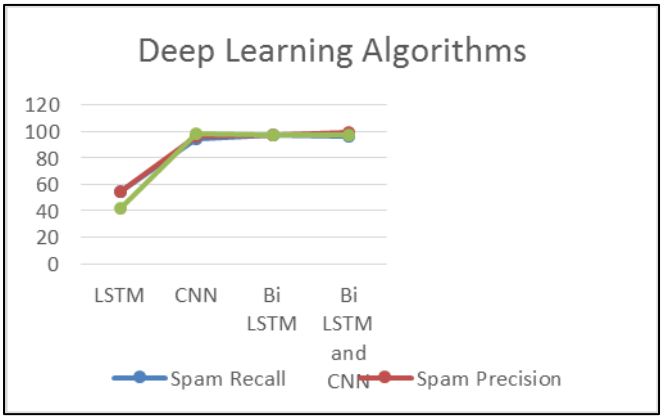
Datasets	SVM	Decision Tree	Naïve bayes	CNN	LSTM	Bi LSTM	Bi LSTM + CNN
Enron	100	94	99	95.7	54.8	98	99
Kaggle	98	96	98	97.4	53.2	97	98.23
Lingspam	100	99	100	98	53.9	96	98.79
Spam Detection	98	97	98	93.1	52.6	98	99
Spam Assassin	99	98	99	96	54.4	99	100
Trec07	99	97	99	98	56.25	96	98.79

Spam recall table

Datasets	SVM	Decision Tree	Naïve bayes	CNN	LSTM	Bi LSTM + CNN	Bi LSTM + CNN
Enron	98	92	99	93.7	49	96	98.7
Kaggle	96	91	98	95.5	52.4	97	95.22
Lingspam	100	96	100	96.2	53.7	96	96.47
Spam Detection	93	89	99	92.7	51.6	97	98.7
Spam Assassin	99	95	99	93.8	53.2	100	95
Trec07	96	92	98	92.7	55.99	96	96.47

Algorithm	Spam recall	Spam precision	Spam Accuracy
Decision tree	92.5	96.83	90.34
Support vector machine	97	99	93.31
Naïve bayes	98.83	98.3	90.34
LSTM	54.64	54.19	42
CNN	94.1	96.36	98.37
Bi LSTM	97	97.33	96.98
Bi LSTM and CNN	96.76	98.96	97.47





V. Conclusion And Future Work

From the results given in the above table 3 we conclude that the hybrid algorithm Bi LSTM + CNN has the best performance based on all the three parameters Spam Precision, Spam recall and Spam Accuracy. However using this algorithm is quite costly. Therefore, we need to find an efficient way to reduce the computational cost in the future.

References

[1] A Proposed Data Science Approach for Email Spam Classification using Machine Learning Techniques, Aakash Atul Alurkar*,Sourabh Bharat Ranade, 978-1-5386-3197-3/17/$31.00 ©2017 IEEE
[2] Identification of Spam Email Based on Information from Email Header, Shukor Bin Abd Razak, Ahmad Fahrulrazie Bin Mohamad, 978-1-4799-3516-1113/$31.00 ©2013 IEEE
[3] Content-Based Spam Detection in Email using Bayesian Classifier, Sunil B. Rathod, Tareek M. Pattewar, 978-1-4799-8081-9/15/$31.00 © 2015 IEEE
[4] Sunday Olusanya Olatunji, "Improved email spam detection model based on support vector machines", Neural Computing and Applications, vol.31, pp. 691-699, 2019
[5] An Overview of RNN and CNN Techniques for Spam Detection in Social Media, Jain et al., International Journal of Advanced Research in Computer Science and Software Engineering 6(10), October - 2016, pp. 126-132
[6] Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism, 2169-3536 2019 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information
[7] Should We Be Concerned with Spam Emails? A Look at Its Impacts and Implications, Yanti Rosmunie Bujang, Husnayati Hussin, 978-1-4799-0136-4/13/$31.00 ©2013 IEEE

