BI-MODAL EMAIL SPAM DETECTION USING RECURRENT AND CONVOLUTION NEURAL NETWORK TECHNIQUES

By

SULAIMON BASHIR *

OLAWALE SURAJUDEEN ADEBAYO **

KEHINDE LAWAL ***

MUHAMMAD ABDULLAHI ****

** Robert Gordon University, Aberdeen, Scotland, United Kingdom.*
*** Department of Computer Science, International Islamic University, Malaysia.*
**** Department of Computer Science, Ladoke Akintola University of Technology, Ogbomoso.*
***** Department of Computer Science, Federal University of Technology, Minna, Nigeria.*

Date Received: 15/05/2023 *Date Revised: 05/06/2023* *Date Accepted: 07/07/2023*

ABSTRACT

The increasing adoption of electronic emails as a means of communication, both at the commercial, government, and individual levels, serves as an impetus for attackers to compromise communication. Consequently, numerous machine learning techniques have been developed for identifying unwanted emails, commonly known as spam. Despite the significant progress reported in existing literature, most studies do not integrate the detection of both textual and image-based spam. In this paper, two deep learning techniques that detect both textual and image-based spam were evaluated. First, the Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) is studied, training them on various text-image features to explore their effectiveness on an improved dataset. Subsequently, in an effort to outsmart current spam detection techniques, a bi-modal architecture capable of detecting textual spam, image spam, and mixed spam is designed. The experimental results in conjunction with existing transfer learning for effective spam detection is provided.

Keywords: Machine Learning, Bi-Modal, Convolution Neural Network, Email Spam.

INTRODUCTION

Spam emails are unsolicited message content shared through emails to several recipients using electronic devices. In most cases, spammers have no prior relationship with the recipients and collect addresses from various sources such as tagged-filled forms, phone books, and spam messages before sending the spam mails for malicious purposes. Over the last decade, email has become inundated with spam content. Image spam is a new spam tactic developed by spammers that embeds a malicious image with a text message in binary format to avoid detection by text-based spam filters.

The widespread availability and increasing use of the internet have facilitated quick and simple types of online transactions as well as various methods of e-communication, the most common of which is emailing. However, sending and receiving emails as a primary mode of communication has become very common (Al-Shboul et al., 2016). Email is almost a requirement for e-transactions. Despite the availability of various types of e-communications, sending and receiving e-mails has maintained its position as the simplest and fastest method of e-communication. It is one of the most widely used, fastest, and most efficient methods of exchanging information. However, due to the widespread use of email, there has been an increase in the number of problems caused by Unsolicited Bulk Email (UBE) messages, also known as spam. Regardless, emails have remained successful in the field of online business transactions and are now required for other forms of online communication (Verma, 2017).

The increased use of email applications and online transactions via email has contributed to a high rate of email spamming, which has been a major problem in the field of computing. Spamming is a rapidly growing type of attack that, along with phishing, worms, and viruses, is the most dangerous threat to email users (Sharaff et al., 2018). According to (Statista, 2019) the global spam volume as a percentage of the overall e-mail traffic, sorted by month and most recently reported period, was approximately 53.5 percent of total e-mail traffic worldwide. In the second quarter of 2018, China was responsible for the majority of unsolicited spam e-mails, accounting for 4.36 percent of global spam frequency.

The increasing rate of emails spam is continuous and alarming. It has created a major problem for service providers, jeopardizing user confidentiality and causing resource loss. Enormous losses are incurred by organizations, spanning from the waste of bandwidth and mail server load to the impact on client profitability due to the time spent identifying and dealing with spam mail senders. Spam messages are used for a variety of attacks and to bridge security measures, not only increasing device correspondence and storage facility loss. This violence has the potential to be used to abuse client data and steal valuable sensitive information such as passwords and financial information. The high volume of spam mail that circulates among networked computers has a negative impact on email server memory space, bandwidth, application terms, and processing power.

Spam email is becoming more of a threat on an annual basis, accounting for more than 77 percent of all email traffic globally (Fonseca et al., 2016). There have been numerous machine learning technique for identifying unwanted spam. Despite the significant progress made in the figures of existing literatures reviewed, no classification technique has achieved 100 percent accuracy (Chopra & Gaikwad, 2015). For classification, each algorithm makes use of a limited set of features and properties. Because of the critical role of separating spam from ham (non-spam) email, two approaches have emerged: Knowledge engineering, which uses rules to separate spam from ham email, and Content base engineering, which uses a heuristic method to separate spam from ham email by learning from incoming mail and training the remaining proportional email so that the process can continue without further training. The majorities of current spam filters are only capable of identifying text based spam or image spam. In this research work, a multi-modal architecture that can identify textual spam, image spam, and mixed spam is developed.

1. Related Work

The rapid rise in email spam filtering is attributed to the rise in spam emails, which has prompted several comparative studies by neuroscientists on the effectiveness of spam picture-based email classification algorithms employing hybridization metrics, which has the interest of the global research community. As a result, determining which technique works best for a specific statistic is crucial in order to assure accurate email classification as spam or not. The machine learning technique has received a lot of attention, and there are numerous algorithms that may be utilized in e-mail filtering. Naive Bayes, support vector machines, Neural Networks, K-nearest neighbor, Rough sets, and the artificial immune system are among them. There are two general ways to spam detection or email filtering at the moment. Knowledge engineering and content-based filtering in e-mail classification are two examples. A set of rules must be included in the knowledge engineering technique to classify emails as spam or ham. A set of such rules should be developed by the user, the filter, or another authority (like the software company that provides a particular rule-based spam filtering tool). Because the rules must be changed and maintained on a regular basis, this technique yields no promising outcomes because it is not a real-time procedure and is inconvenient for most users (Foqaha, 2016).

Abdullahi et al. (2021) devised a performance analysis-based technique for email spam detection using classification techniques such as Bayesian logistic regression in their research on the quantitative comparison of classification techniques for email spam detection. Some of the concepts utilized in this research include naive Bayes, logic boost, rotation forest, neural network, logistic model tree, rep tree, Support Vector network, voted recurrent neural networks, lazy Bayesian rule, multilayer perceptron, random tree, and J48. Using the spam base dataset and weka data mining tools, the accuracy, precision, recall, F l-Measure, root mean squared error, receiver operator characteristics area, and root relative squared error of these strategies were categorised. The rotating forest algorithms were known to have the highest F1-measure, whereas the naive bayes methods had the lowest F1-measure. The Receiver Operating Characteristic (ROC) curves are utilized for randomly selected positive and negative instances, alongside the rotating forest algorithm, to determine the likelihood. The random tree received the lowest score of 0.90, while the ROC curves received the best score of 0.98. The researchers also utilized kappa statistics to derive the statistical results, demonstrating that the rotating forest algorithm yielded significantly superior outcomes, achieving an approximate accuracy of 87.9%. The article found that rotation forest classifiers have the highest accuracy of 94.2 percent, followed by J48 (92.3 percent), naïve bayes (88.5 percent), and multilayer perception (93.2 percent).

Using Python Machine Learning Techniques (Mohammed et al., 2013), proposed a new approach for Classifying Region's Largest Email (UBE) Using Python Machine Learning Techniques with Spam Filtering, which actually achieves the task by creating a spam ham dictionary from the given training data and applying a data mining algorithm to filter the training and testing data. After applying various classifiers on a 1431 dataset, the approach predicts that. Naive Bays and SVM classifiers are the most commonly used spam classifiers.

Singh and Bhardwaj (2018) reviewed the solution and classification process of spam filtering and proposed a combining classification strategy to produce a better spam filtering outcome in their research on spam email detection utilizing classification techniques and global training sets. Data mining was employed to compile all of the information about previous spam filtering failures, successes, and present issues. A binary value system was used in which 1 represents spam email and 0 represents non-spam (ham) email. However, the success rate was low, so NB, KNN, SVM, and Artificial Neural Network classification methods were used to determine accuracy. Subsequently, a classification strategy for spam filtering was adopted based on the two methodologies of machine learning and knowledge engineering effectiveness. Data was collected from a user training set, spam emails were compared and discovered, and the categorization algorithm was then optimized using a global training set. The precision rate is increased by at least 2% when this technique is used.

In their work on an approach for malicious spam detection in emails, with a comparison of different classifiers, Sah and Parmar (2017) proposed a procedure of techniques for detecting malicious spam. This involved feature selection and improving the training accuracy and time for malicious spam detection systems. The performance of Naive Bayes (NB) and Support Vector Machine (SVM) classifiers was compared based on calculation time and accuracy reports. Preparing text data, building a word dictionary, extracting features, and training the classifier were all phases in this novel approach.

Yuksel et al. (2017) designed a machine learning-based predictive analytics system for the spam problem. For spam filtering, the authors used the support vector machine and decision tree. The decision tree was utilized in data mining, while support vector machines were employed as a supervised learning model to evaluate data for spam classification. The data was first divided into two sections: one for training and the other for testing. After that, the algorithm was trained and evaluated using the Microsoft Azure platform, which provides machine learning capabilities, and the outcomes of the support vector machine and the decision tree method were compared side by side. The SVM scored 97.6% after the analysis, compared to 82.6 percent for decision tree findings. The SVM classifier outperformed decision tree classifiers, according to these estimates.

Swapna et al. (2017) investigate spam email detection using supervised machine learning algorithms. Inductive or supervised machine learning techniques were used by the authors. The algorithms in this method learn from a training dataset that contains both inputs and outputs results, resulting in the creation of a new model. After that, the new model is put to the test on new categorization samples. In the case of binary categorization, the output falls into one of two categories: spam or ham (that is a legitimate mail). The authors then employed machine learning techniques such as neural networks, naive bayes, support vector machines, lazy algorithms, decision trees, and artificial-immune systems to learn about incoming email behaviors and then classify them as spam or ham email based on their email datasets. During the review of this paper, a brief summary of the various methodologies was examined, and the various performance measures were assessed using the measure of confusion matrix. The neural network was discovered to outperform the rest of the performance measures obtained.

Looking at some of the most common spam traits, trends, and evasion strategies used by spammers, this highlights interesting research approaches as well as some research gaps. The authors claim that filtering e-mail spam is challenging due to the dynamic behavioral character of spam. Therefore, advocating thorough research into spam behavior is essential to better understand its nature and evolution. This understanding can help in developing appropriate anti-spam countermeasures. A taxonomy of content-based email spam filtering, as well as a qualitative review of significant spam email surveys from 2004 to 2015 was carried out. After that, a report on new suggestions and findings of future investigations into machine learning strategies for emerging spam varieties was completed. The author next went over email corpus preprocessing, feature extraction, feature selection, and header and non-content feature analysis. The overview of the various spam filtering strategies utilized prior to machine learning was then spelled out, and the machine learning algorithm application to textual and multimedia content of spam emails was then adequately figured out. Recent methodologies that have emerged, as well as their conventional evaluation metrics, were given special consideration.

The authors of the paper, continuing to improve Email Spam Detection Using a Content-Based Feature Engineering Approach (Wadi et al., 2017), created a comprehensive collection of spam email features representing various interests. They employed a powerful and comprehensive feature extraction tool specifically designed for processing multiple email corpora. The dataset was then utilized to train and evaluate different classification methods. This graph compares the performance of four prominent classifiers when trained with all the features presented in this paper and the results obtained. To assess the created spam detection mode1, the authors used three evaluation measures: Accuracy rate, Precision, and Recall to achieve the results, with the Random forest producing the best classification results in the study.

Chopra and Gaikwad (2015) used a two-stage strategy to classify the textual section of a picture in order to determine whether the words in the letter were spam or not. The researchers noted in their study titled "The Image and text spam filtering" that spammers have devised a new approach to embed spam email inside the image attached to the package, and that OCR tool and Bayesian algorithm were utilized in the initial step. The researchers are led to offer the strategy in an attempt to solve this difficulty. Based on the hybridization of KNN and SVM, a method was proposed. The basic idea is to classify a verification problem's nearby neighbors and produce a close by SVM for the task of separation on the closed array. Their work experiment was carried out using the Dredze dataset and a public dataset, and the results suggest that the results have improved to around 98 percent, but just for accuracy as a performance indicator.

2. Methodology

The Supervised Machine Learning system has two modes of operation, training and testing. In training mode, labeled data is provided to the machine learning system from a training data set. A huge number of emails have been classified as spam or non-spam in this study's labeled training data (that is, ham). The classifier (the part of the machine learning system that does the actually predictions of future email labels) learns from the training data by determining the links between an email's properties and its label.

While in testing mode, the machine learning system is fed with unlabeled data. This data in the case is emails that have not been labeled as either spam or ham. Based on the email's properties, the classifier determines whether it is Spam or Ham. Accessing and structuring the raw email data set, then preparing the data, then examining the data for analysis (EDA), and finally interpreting the performance techniques to choose from. A multi-modal approach for spam detection is provided, aligning with the theme of multi-modal spam detection in the study. To extract textual semantic relational characteristics and generate a classification probability value for the text component of an email as spam, an LSTM model, a type of RNN is utilized. For the image section of the same email, a CNN model is developed to derive the classification probability value of the image segment as spam. These two models produce two distinct classification outcomes. Figure 1 shows the overall framework of the multi-modal architecture as well as the basic procedures for calculating the classification probability value of an email as spam.

2.1 Text Classification Model

In this phase of work, three architectures were designed and analysed, LSTM, CNN, and a hybrid of CNN and LSTM.

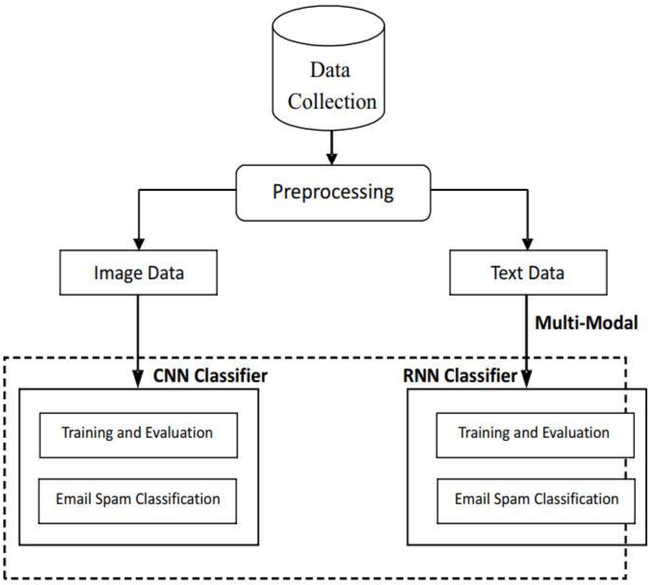

Figure 1. The Architectural Structure of the Model

The LSTM model is made up of one word embedded layer, two LSTM layers, and one Fully Connected (FC) layer.

To get a word vector representation of an email, first capture its text data with the preprocessing technique, then utilize the word embedding technique. After that, the two LSTM layers that have been set to extract features from the text input automatically were employed. Finally, the FC layer and the Softmax activation function to determine whether the text data is spam or ham is used. The optimal values for the five hyperparameters of the LSTM model, learning rate, batch size, epochs, dropout rate, and optimization procedure are determined using the grid search optimization technique.

Let T stand for the email's text data. Input T into the embedding layer to produce a word vector x. x = (xl, x2, …, xl), where xi Rn denotes the n-dimensional word vectors for both the i-th word in the document T as well as matrix xRlxn represents the document T, while l is the maximum size of it and l 500. The memory ct and the hidden state ht are updated at time-step t.

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ C_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} W.[h_{t-1}, x_t], \qquad (1)$$

$$C_t = f_t \bullet c_{t-1} + i_t \bullet c_t, \qquad (2)$$

$$h_t = o_t \bullet \tanh(c_t), \qquad (3)$$

where xt is the current time-input, step's I f, and o are the input gate, forget gate, and output gate activation, respectively, and t is the current cell state, represents the logistic sigmoid function, and means element-wise multiplication. The LSTM model has been trained and optimized.

Algorithm 1: Text Spam Classification Algorithm.

Input: Text Document T

Output: Text spam classification probability value e

- Input T into the word2vec toolkit to get the word vector x, x = (x1, x2,· · · , xl).
- For the first LSTM layer, input x at time t and complete the following calculations:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ C_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} W.[h_{t-1}, x_t],$$

$$C_t = f_t \bullet c_{t-1} + i_t \bullet c_t, \qquad (4)$$

$$h_t = o_t \bullet \tanh(c_t),$$

- The text feature vector h = (h1, h2, , h64) is obtained by the first LSTM layer.
- For the second LSTM layer, input h at time t.

  Finally, k = (k1, k2, k32) to obtain a more abstract text feature vector k.
- Apply the Softmax activation function to the FC layer to obtain the text classification probability value e;
- return e;

2.2 Image Classification Model

A CNN model is developed to classify emails in this portion of our research. The hyperparameters of the CNN model, the CNN architectures, and the architectures of the designed CNN image-based classifiers, along with the range and optimal values of these hyperparameters selected by the CNN model, have all been implemented. A CNN model is developed to classify emails in this portion of our research.

Algorithm 2: Algorithm for Image Spam Classification.

Input: Image m, size 128 x 128 RGB.

Output: The probability value for the image spam g.

- Input m to the three convolutional layers, and you'll obtain d, d= (d1, d2, d64);
- Input d to the first two FC layers to obtain feature vector c, where c = (c1, c2, c32);
- To acquire the probability value of categorization g, input c to the last FC layer, which contains two neurons, using a Softmax activation function.
- Return g;

2.3 LSTM-CNN (Multi-Modal)

The goal is to obtain the most accurate spam classification probability value for an email by combining the classification probability value of an email text part with the classification probability value of the same email image component.

- Combining the two classification probability values of the LSTM and CNN models q, q∈R1×4 ; to create a feature vector.
- Inputting q into the FC layer to generate a comprehensive feature vector.
- Inputting the comprehensive feature vector into the logistic layer, which is made up of two neurons and uses the logistic regression function as an activation function to get the most accurate classification probability value of the email as spam.
- Assume the classification probability dataset for the generated model is as follows:

  $D=\{(q_1, y_1), (q_2, y_2), (q_v, y_v)\}, q_i \in R^{l\times 4}, y_i \in \{0,1\}$ (5)

The logistic regression function's conditional probability distribution is as follows:

$$P(Y = 1/q) = p(q) = \frac{e^{-\ w^T.q}}{1+ e^{-\ w^T.q}} \qquad (6)$$

$$P(Y = 0/q) = 1- p(q) = \frac{1}{1+ e^{-\ w^T.q}} \qquad (7)$$

As the loss function, the log-likelihood function is used, which has the following formula:

$$\begin{aligned} L(w) &= \sum_{i=1}^{u} [y_i \log p(q_i) + (1-\ y_i)\log(1-\ p(q_i))] \\ &= \sum_{i=1}^{u} [y_i \log \frac{p(q_i)}{1-\ p(q_i)} + \log(1-\ p(q_i))] \qquad (8) \\ &= \sum_{i=1}^{u} [y_i(w.q_i) -\ \log(1+ e^{(w.q_i)})] \end{aligned}$$

The Adam optimization algorithm determines the optimal value of L (w). Furthermore, by optimizing L, the ideal estimate value of the parameter w can be derived (w). If p > 0.5, the email is spam, else it's just a regular email.

3. Dataset

In this research, two publically available datasets that include text and photos were employed. All of these datasets contain spam and ham content derived from real email.

3.1 Image Spam Hunter (ISH) Dataset

Image Spam Hunter's creators gathered a big sample of image spam and a similarly large sample of ham images. This information is referred to as the ISH dataset. Following data cleaning, 920 spam photos and 810 ham images from the ISH dataset were preserved for this study. Table 1 shows the Image Spam Hunter Email Dataset Distribution.

3.2 Enron Spam Dataset

The Enron dataset consists primarily of emails exchanged by senior management of Enron Corporation (Klimt & Yang, 2004). This is the second dataset used, and it comprises a publicly available dataset obtained from the Enron corpus. Only 20,000 text emails for both spam and ham in an equal ratio is selected after discarding duplicates and randomly choosing from the pool of 33,645 text emails. Among these, 17,108 emails are labeled as ham, and 16,537 emails are labeled as spam. Table 2 shows the Enron Email Dataset Distribution. The testing of 30% to 70% of training was carried out for all the datasets respectively.

4. Performance Evaluation

Various performance indicators, such as accuracy, recall, precision, and F1-score, were employed to assess the usefulness of the suggested strategy. Furthermore, The Confusion Matrix, often known as the Error Matrix, is a popular tool for analyzing performance algorithm results. The confusion matrix is used as the primary evaluation metric for spam detection.

Email Dataset	SPAM	HAM	Total
ISH Dataset Distribution	879	810	1689
Percentage Distribution (%)	50.8	49.2	100
Training Distribution (70%)	601	582	1182
Testing Distribution (30%)	258	249	507

Table 1. Image Spam Hunter Email Dataset Distribution

Email Dataset	SPAM	HAM	Total
Enron Dataset Distribution	1000	10000	20000
Percentage Distribution (%)	50	50	100
Training Distribution (70%)	7000	7000	14000
Testing Distribution (30%)	3000	30000	6000

Table 2. Enron Email Dataset Distribution

4.1 Confusion Matrix (CM)

A Confusion Matrix is a table that is employed frequently to describe how well a classification model (classifier) performs on a set of test data with known true values. It is a brief and logical categorization of task and prediction outcomes. The confusion matrix is shown in Table 3, along with the anticipated column and actual class row names.

The following components from the definition of Confusion Matrix (CM) is deduced.

- *True Positive (TP):* The percentage of spam classified correctly.
- The number of correctly categorized legitimate emails in the True Negative (TN) category (Ham).
- The number of genuine emails that are incorrectly labeled as false positives (Ham) (FP).
- The number of spam messages that have been misclassified as false negatives (FN).

Each of the metrics described above is applied to data in order to evaluate the performance measure and compare the various classification reports in order to determine the best classifier model efficiency of the algorithm in the supervised machine learning method being used. A good model must be able to generalize successfully to test data that is significantly different from the training data. On test data, a model built on training data by learning which scenarios fill the best may not perform well. The confusion matrix is also calculated using the following parameters.

4.1.1 Accuracy

The relationship between classification rate and accuracy is:

$$AccuracyRate(Acc) = \frac{TP + TN}{TP + TN + FP + FN} \qquad (9)$$

This metric indicates how well a binary classification test works. What percentage of correctly predicted events actually occur. Accuracy alone isn't a good measure because it doesn't tell you how effectively the model

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Table 3. A Confusion Matrix

detects positives and negatives independently. It assumes that the costs of both types of errors are equal. Depending on the difficulty, 99 percent accuracy might be outstanding, decent, middling, poor, or even terrible.

4.1.2 Recall

$$Recall = \frac{TP}{TP + FN} \qquad (10)$$

The recall is calculated as the ratio of correctly classified positive spams to total positive spams. It describes how effective a test is at detecting positive spam. To put it another way, positive outcomes are predicted to be positive. The class was accurately identified if the recall was high (small number of FN). A good model should have a high recall rate. Sensitivity or TP Rate are other terms for recall. The recall (Rs) metric indicates how many spam messages the filter successfully prevented (i.e. the degree of its effectiveness of blocking actual spam correctly).

4.1.3 Precision

$$Precision = \frac{TP}{TP + FP} \qquad (11)$$

The total number of properly categorized positive spam is divided by the total number of expected positive spam to determine precision rates. A high precision suggests that a test data email that has been flagged as spam is, in fact, spam (small number of FP Rate). It is the percentage of positive projections that are correct. A high level of precision is excellent. Precision (Ps) measures the proportion of messages labeled as spam by the filter that were, in fact, spam.

4.1.4 False Positive Rate

$$FPRate = \frac{FP}{FP + TN} \qquad (12)$$

This metric describes how well a model detects negatives. A high FP Rate is caused by a model that forecasts as positive when it is actually negative. This metric is sometimes given a one-star rating for specificity, which is defined as (TN Rare).

$$Specifity = \frac{TN}{TP + FP} \qquad (13)$$

It's preferable to have a high specificity (the ability to correctly forecast all negatives). The bulk of positive spam is successfully identified (low FN), but there are a lot of false positives. This shows that a lot of positive spam (high FN)is missed, but those were expected to be positive are indeed positive (low FP).

4.1.5 FI-Measure

Because there are two measurements (precision and recall), having a better measurement that encompasses both is advantageous. The F1-measure is calculated using Harmonic Mean rather than Arithmetic Mean since extreme values are penalized more severely. The smaller precision or recall value will always be closer to the F1-Measure. It is highly beneficial to have a high F1-measure.

$$F1 - Measure = \frac{2 * Recall * Precision}{Recall + Precision} \qquad (14)$$

5. Results and Discussion

The evaluation results for both text and image spam classification is given in this subsection, as well as some analysis and discussion of the experimental data. This is for multiple fold cross validation to evaluate the model performance. Table 4 shows the values of the measurement metrics that were used. Figure 2 shows the model performance on text and image data.

To further validate the model's performance, the comparison is made to a large set of well-performed models using the same or different datasets. The text dataset is used to test Character-level CNN (Char-CNN), BiLSTM (Bi-directional LSTM), Naive Bayes, and the Immune

Data Type	Accuracy	Recall	F1-Score	Precision
Text Dataset	0.98	0.96	0.97	0.96
Image Dataset	0.98	0.97	0.97	0.96

Table 4. Model Performance for Text and Image Data

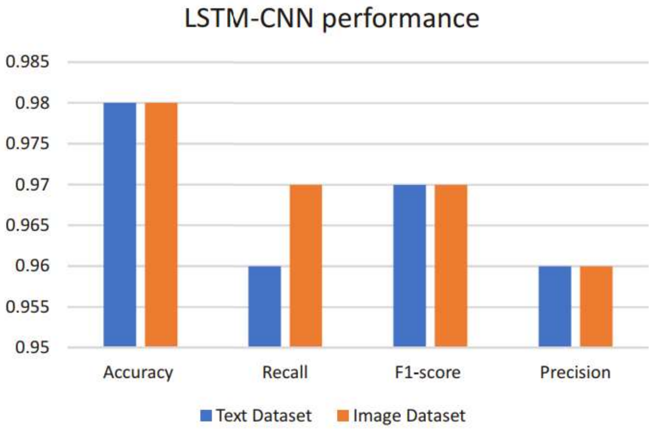

Figure 2. Model Performance on Text and Image Data

Cross-Regulation Model (ICRM). ME, LSVM, CNN, Naive Bayes, and ID3 Decision Tree are compared on the image dataset. The grid search optimization approach was used to pick the best hyperparameters for the four models to show that the LSTM CNN model is superior. The best hyperparameters of the SVM model through the grid search optimization algorithm, for example, are l and 0.00l, respectively, whereas the optimal hyperparameter K of the traditional k-NN model is l. Against comparing our designed model's performance to a huge number of well-performing models utilizing the same or alternative datasets, Tables 5 and 6 show the performance of existing state-of-the-art text and image spam categorization models. Table 7 shows the model performance for text and image data. Figure 3 and Figure 4 show the performance comparison of the models on text datasets. According to Table 7, the LSTM-CNN model outperformed the other models on the text and image datasets.

Year	Dataset	Model	Accuracy	Precision	Re-Call	F- Measure
2016	Enron	Char-CNN	0.96	-	-	-
2018	Enron	BiLSTM	0.964	-	-	-
2015	Enron	Naïve Bays	0.96	-	0.960	-
2017	Enron	ICRM	0.94	-	0.94	-
2019	SMS Spam	LSTM	0.91%	-	0.90	0.90
2016	Enron	Decision	96%	98%	94%	-

Table 5. The Performance of Existing Text Spam Classification Models in their Current State

Year	Dataset	Techniques	Accuracy	Precision	Re-Call	F- Measure
2015	Dredze	SVM	90%	-	-	-
2015	Spam Base	Naïve	84%	89%	78%	-
2015	ISH Dataset	KNN	93/74	97/96	91/01	94/35
		Naïve	99/19	98/50	98/52	99/25
2018	Spam base	Random	94.2%	94.2%	94.2%	94%
		Naïve	88.2%	88.5%	88.5%	88.5%
		Multilayer	93.2%	93.3%	93.2%	93%
		J48	92.3%	92.3%	92.3%	92.3%
2017	Dredze	Naïve	98%	-	-	-
2015	Spam base	SVM	79.50%	79.02%	68.69%	-
		Naïve	76.24%	70.59%	72.05%	-
2018	Spam base	ANN	92.41%	92.40%	92.4%	-
2018	ISH	Naïve Bayes	0.85	-	0.91	-
2016	Dredze	SVM	0.97	0.97	0.68	-
2018	ISH	J48	0.92	0.92	0.92	0.92
2017	Spam base	ID3-DT	0.89	-	0.90	-

Table 6. The Current Performance State of Image Spam Classification Models

Data Type	Model	Accuracy	Recall	F1-Score	Precision
Text Dataset	Char-CNN	0.95	-	-	-
	ICRM	0.94	-	0.95	-
	Naïve Bays	0.96	-	0.96	-
	BiLSTM	0.95	-	-	-
	DT	0.96	0.98	0.94	-
	LSTM-CNN	0.98	0.96	0.97	0.96
Image Dataset	Naïve Bayes	0.85	-	0.91	-
	SVM	0.97	0.97	0.68	-
	Naïve Bayes	0.96	-	0.96	-
	J48	0.92	0.92	0.92	0.92
	ID3 D.T	0.89	-	0.90	-
	LSTM-CNN	0.98	0.97	0.97	0.96

Table 7. Model Performance for Text and Image Data

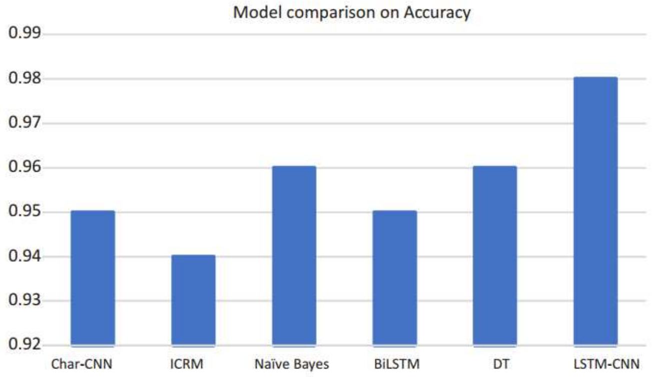

Figure 3. Performance Comparison of the Models on Text Datasets

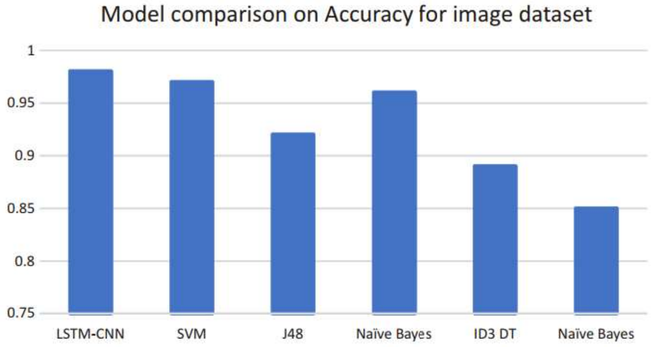

Figure 4. Performance Comparison of the Models on Image Datasets

Conclusion

In order to improve spam detection rate, a multi-modal architecture based on model fusion is first introduced by combining the Convolutional Neural Network (CNN) and the Long Short-Term Memory (LSTM) networks via the logistic regression method to implement spam detection in a variety of email formats. The new model has the advantage of being able to filter hybrid spam as well as spam including either text or image data, whereas prior models can only handle text-based or image-based spam. There have been numerous machine learning technique for identifying unwanted email spam. Despite the significant progress made in the figures of existing literatures reviewed, no classification technique has achieved 100 percent accuracy for classification, each algorithm makes use of a limited set of features and properties. As a result, determining the best algorithm is a critical task because their strengths must be balanced against their limitations.

Significant progress has been made, based on the volume and figures of literature reviewed; however, more research is needed to improve the performance of multi-modal systems on artificial immune systems, as well as to focus on the availability of well-labeled datasets to enable successful spam filtering. There has also been a rise in the use and application of the internet, which has been linked to the rise of spam images.

In this research, a variety of real-world text and image spam datasets were leveraged to develop powerful classifiers based on deep neural networks and convolutional neural networks. This work was able to produce better outcomes when compared to the previous results. These techniques were able to learn even with the increased dataset offered. However, despite the fact that it worked better, there is still potential for development.

References

[1]. **Abdullahi, M., Mohammed, A. D., Bashir, S. A., & Abisoye, O. O. (2021, February).** A Review on Machine Learning Techniques for Image Based Spam Emails Detection. In 2020 *IEEE 2nd International Conference on Cyberspac (CYBER NIGERIA)* (pp. 59-65). IEEE. https://doi.org/10.1109/CYBERNIGERIA51635.2021.9428826

[2]. **Al-Shboul, B., Hakh, H., Faris, H., Aljarah, I., & Alsawalqah, H. (2016).** Voting-based classification for E-mail spam detection. *Journal of ICT Research & Applications*, 10(1), 29-42.

[3]. **Chopra, N. D., & Gaikwad, K. P. (2015).** Image and text spam mail filtering. *International Journal of Computer and Electrical Engineering (IJCTEE)*, 5(3), 71-105.

[4]. **Faris, H., Alqatawna, J. F., Ala'M, A. Z., & Aljarah, I. (2017, October).** Improving email spam detection using content based feature engineering approach. In 2017 *IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)* (pp. 1-6). IEEE. https://doi.org/10.1109/AEECT.2017.8257764

[5]. **Fonseca, O., Fazzion, E., Cunha, I., Las-Casas, P. H. B., Guedes, D., Meira, W., ... & Chaves, M. H. (2016).** Measuring, characterizing, and avoiding spam traffic costs. *IEEE Internet Computing*, 20(4), 16-24. https://doi.org/10.1109/MIC.2016.53

[6]. Foqaha, M. A. M. (2016). Email spam classification using hybrid approach of RBF neural network and particle swarm optimization. *International Journal of Network Security & Its Applications*, 8(4), 17-28.

[7]. Klimt, B., & Yang, Y. (2004, September). The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning* (pp. 217-226). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-30115-8_22

[8]. Mohammed, S., Mohammed, O., Fiaidhi, J., Fong, S., & Kim, T. H. (2013). Classifying unsolicited bulk email (UBE) using python machine learning techniques. *International Journal of Hybrid Information Technology*, 6(1), 43-56.

[9]. Sah, U. K., & Parmar, N. (2017). An approach for malicious spam detection in email with comparison of different classifiers. *International Research Journal of Engineering and Technology (IRJET)*, 4(8), 2238-2242.

[10]. Singh, V. K., & Bhardwaj, S. (2018). Spam mail detection using classification techniques and global training set. In *Intelligent Computing and Information and Communication: Proceedings of 2nd International Conference (ICICC)* (pp. 623-632). Springer Singapore.

[11]. Statista. (2019). *Monthly Share of Spam in the Total E-Mail Traffic Worldwide from January 2014 to December 2022*. Retrieved from https://www.statista.com/statistics/420391/spam-email-traffic-share/

[12]. Torabi, Z. S., Nadimi-Shahraki, M. H., & Nabiollahi, A. (2015). Efficient support vector machines for spam detection: A survey. *International Journal of Computer Science and Information Security*, 13(1), (pp. 11).

[13]. Verma, T. (2017). E-Mail spam detection and classification using SVM and feature Extraction. *International Journal of Advance Research, Ideas and Innovations in Technology*, 3(2), 1491-1495.

[14]. Verma, T. (2017). E-Mail spam detection and classification using SVM and feature Extraction. *International Journal of Advance Research, Ideas and Innovations in Technology*, 3(2), 1491-1495.

[15]. Wang, J., & Katagishi, K. (2014). Image Content-Based "Email Spam Image" Filtering. *Journal of Advances in Computer Networks*, 2(2), 110-114. https://doi.org/10.7763/JACN.2014.V2.92

[16]. Yüksel, A., Çankaya, Ş., & Üncü, I. (2017). Design of a machine learning based predictive analytics system for spam problem. *Acta Physica Polonica A*, 132(3), 500-504.

ABOUT THE AUTHORS

Sulaimon A. Bashir obtained his Doctorate degree in Computing from Robert Gordon University, Aberdeen, Scotland, United Kingdom. He was a recipient of the Ph.D scholarship from the National Information Technology Development Agency. His research interests include the application of Machine Learning to Computer Vision, Natural Language Processing, and Cybersecurity. He has published several research papers in peer-reviewed journals and conferences related to the aforementioned research areas.

Olawale Surajudeen Adebayo earned his Ph.D. in Computer Science from the International Islamic University, Malaysia. His research interests include Computer and Information Security, Cryptography, Malware Detection, Data Mining, and IT Entrepreneurship. He has published and presented many peer-reviewed papers in reputable and renowned international journals or conferences.

Kehinde Lawal obtained his Master's degree in Computer Science from Ladoke Akintola University of Technology, Ogbomoso. He has worked as a Principal Programmer at the Information Technology Services Department, FUT Minna, where he developed various web-based and desktop-based applications for the smooth running of university services.

Muhammad Abdullahi obtained his Master's degree in Computer Science from the Federal University of Technology, Minna, Nigeria. He is currently focusing on research in Computer Vision System Security.