Email Spam Detection Using LSTM with Attention Mechanism

Allam Bala Praneeth reddy[1], Karrepu Hema Charan Reddy[2], Marujolla Nihar Reddy[3]

[1]2310030431@klh.edu.in, [2]2310030419@klh.edu.in, [3]2310030441@klh.edu.in

[1]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India.

[2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India.

[3]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India.

Abstract: Spam email detection is an important domain of cybersecurity because spam emails usually contain phishing URLs, malicious attachments, or fraudulent advertising content. These emails not only interfere with communication but also lead to major threats like data robbery, money loss, and system compromise. The conventional rule-based spam filters have been ineffective in dealing with these issues as they lack high adaptability in handling changing spam tactics. In this research, we propose a deep learning-based solution for spam filtering with an LSTM network with an attention mechanism. The model is formulated to efficiently capture sequential and contextual information from email content through highlighting key words—like "urgent," "free," or "win"—that frequently occur in spam mail. A conventional dataset of pre-labelled spam and non-spam (ham) emails was utilized for training and testing. To preprocess the data, preliminary steps like tokenization, padding, and vectorization were involved. The architecture of the model involves an embedding layer, a temporal dependency-learning LSTM layer, and a dedicated attention layer for identifying contextually important words. The ultimate classification output is obtained via a fully connected layer with a sigmoid activation function. Experimental outcomes show that the attention mechanism LSTM greatly surpasses standard Algorithms utilizing machine learning on accuracy, precision, recall, and F1-score. The model also enables real-time classification, enabling users to provide emails and receive instant predictions. This paper establishes the merits of using LSTM in conjunction with attention mechanisms for resilient and adaptive detection of spam email, and paves the way for future improvement with transformer models or multilingual data.

Keywords

Spam detection, Email Spam Filtering, LSTM with Attention, Machine Learning, Email Classification, Cybersecurity.

1. Introduction

Spam email detection is an essential cybersecurity research area, essential in averting phishing, malware, and scams. The conventional rule-based approaches are falling out of Favor, to be replaced with more flexible machine learning methods more able to contend with the continuously evolving tactics of spammers. Models of machine learning such as Naïve Bayes and Support Vector Machines (SVM), and Artificial Neural Networks (ANN) have been deeply studied for spam email classification. These models are appreciated due to their capability of detecting email patterns, which usually

perform better than conventional ones. The efficiency of these models greatly relies on the quality of feature selection and pre-processing of data. Hybrid models that integrate several machine learning techniques usually perform better than single models, as they can detect a large variety of patterns in the data.

Dimensionality reduction has been proposed to enhance the efficiency for the detection of spam. With the features of the dataset being kept small, these models are capable of classifying quickly while maintaining the classification accuracy in place. Having the email metadata like the information of the sender and email header data included also adds to the accuracy because it provides additional information to the classifier upon which to base its decision. Also, probabilistic methods including Bayesian networks and ensemble methods have been employed to improve detection in cases of noisy and uncertain data. However, detection of sophisticated spam emails that are very close to real emails remains an issue, and therefore deep learning paradigms for developing improved spam detection systems have gained immense attention.

Advanced neural network techniques like Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (Bi-LSTM) are recently been used in research to identify sequential dependencies and contextual meaning in email messages. If they are used in tandem with attention mechanisms, they have the potential to enhance precision and recall. Research has proven the effectiveness of using word embeddings with LSTM as well as merging feature extraction with BiLSTM to identify wide-ranging spam structures, a distinct leap in research.

These developments highlight the increasing significance of incorporating sophisticated deep learning models in adaptive spam detection systems, with a focus on sophisticated architectures such as BiLSTM with attention mechanisms to help fight increasingly sophisticated spam strategies.

2. Literature Review

Spam email detection is an important research area since it is a central point in preventing cyber-attacks such as phishing, malware, and spamming. Various machine learning methods have been outlined as an alternative to rule-based systems traditionally used in spam detection. These approaches are based on the capability of algorithms to identify trends from big data and to adapt in response to evolving spam techniques [1]. Algorithms in machine learning, including Naïve Bayes, support vector machines (SVM), and artificial neural networks (ANN) have been extensively researched for spam classification. These algorithms are advantageous since they can classify emails using learned patterns, usually performing better than the traditional approaches. Feature selection and pre-processing are important in a bid to achieve the best performance of these classifiers. Hybrid models combining more than one machine learning approach have been found to be better than a single model since they can detect a more comprehensive set of patterns in the data [2].

Dimensionality reduction has been proposed at times as a method of improving the efficiency of spam detection. Reduction of the size of the feature in the dataset can cause models to compute faster, without compromising much classification accuracy. Addition of email metadata, including the sender details and the email header, increases the accuracy of detection. Such an addition provides the classifier with more context information to choose the better class for prediction [3]. Other studies have looked at the use of ensemble techniques and probabilistic models in an attempt to improve the accuracy of spam classification. Ensemble techniques provide predictions based on more than one model, which tends to result in more stable classification. Probabilistic models, including Bayesian networks, are very useful to use in order to manage noisy uncertain data in real and representative real-world datasets for emails [4]. In recent studies, there has also been interest with the use of Spam filtering systems that are capable of adjusting to the changing nature of spam techniques. This includes adaptive models, which learn from the new patterns in spam emails. Despite these developments, there are still problems with the detection of smart spams when they are as good as communication. There is therefore increased interest in the implementation of deep learning models in order to improve the accuracy and stability of spam email detection [5].

Some of the recent advances in spam filtering have been rooted in the use of advanced neural network architectures, specifically Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (Bi-LSTM) are models because these models are capable of learning sequential relationships and contextual meaning in text data. AbdulNabi and Yaseen [6] demonstrated how word embeddings and LSTM networks significantly improved spam email classification by learning sophisticated word relationships. Moreover, Ragesh et al. [7] proposed a hybrid BiGRU-A-LSTM model with attention that emphasized effective words in spam filtering—resulting in improved precision and recall.

Ullah and Rahman [8] used a CNN-BiLSTM-based model to extract local features and determine long-term dependencies first before proving stronger ability to detect different spam structures. Tusher et al. [9] proved the strength of using word embeddings with LSTM layers, citing the strength of using them in multilingual spam datasets. An SCIRP paper [10] also explored the use of attention-based deep neural networks in spam classification and observed that the use of attention-based models enhanced interpretability as well as classification strength via the dynamic weighing of spam-suggestive words like "win," "cash," and "offer."

These findings collectively point towards the growing demand for attention-based deep learning models in adaptive anti-spam filters. They also validate the use of more advanced models like BiLSTM and hybrid attention networks to combat adaptive spam strategies effectively.

3. Methodology

In our work, we introduce a deep learning-driven spam detection model for emails that takes advantage of the strength of Long Short-Term Memory (LSTM) networks enhanced with an attention mechanism. The aim here is to enhance the classification performance by allowing the model to acquire knowledge of the sequence structure and context importance of words in an email. The dataset consisted of pre-classified email messages as either spam or ham. First, data preprocessing was carried out by substituting missing values and converting the labels of categories into binary form (spam = 0 and ham = 1). The Tokenizer tool of Keras was utilized to convert textual data into sequences of integers, with the vocabulary restricted to the 10,000 most common words and an out-of-vocabulary (OOV) token assigned to any unknown terms. These sequences were padded to a constant size of 100 tokens to obtain a consistent input size for batch processing during training.

The code starts by loading and preparing the dataset using pandas and NumPy libraries. Messages were cleaned, tokenized, and transformed into padded sequences. Following preprocessing, the data were partitioned into training and evaluation subsets in an 80:20 split in order for the model to be tested on previously unseen data. The model architecture was constructed using Keras's Functional API. The initial layer is a layer for embedding that converts word indices to dense 128-dimensional vectors so that the model can capture the semantic connections among words. This embedding output is then fed into an LSTM layer of size 128, which is appropriate for capturing the sequential dependencies within the text.

The unique aspect of the model is the custom attention layer, which is designed in the capacity of standalone function within the code. This layer first performs a dense transformation to produce attention scores, followed by a SoftMax operation for normalization over all time steps. The attention mechanism then actually draws emphasis on the most critical words of a message—e.g., "win," "free," or "urgent"—which are frequently robust indicators of spam. The attention scores are used to weigh the outputs of the LSTM, and the summation is computed to acquire a representation of fixed length of the message. It was passed via a dense layer utilizing sigmoid activation to generate the ultimate classification into two categories output.

The framework was trained using the binary logarithmic loss function and optimized using the Adam optimizer. If there were previously trained weights, they were loaded; otherwise, training of the model was performed for 10 epochs, and the weights were saved for later use. Model evaluation is performed not only by accuracy, but also by precision, recall, and F1-score. These measurements are calculated from scikit-learn's integrated evaluation tools and output for the training

and testing sets. Precision and recall are particularly important in spam filtering, where avoiding false positives (non-spam marked as spam) and false negatives (slipping spam past undetected) are important. Finally, the code has a user-interactive loop whereby users can enter an email message and obtain real-time feedback as to whether the email is spam. This is achieved by tokenizing the user input, padding it, and using a trained model to predict it.

In comparison to the conventional learning algorithms like Naïve Bayes, Decision Trees, and Support Vector Machines, as presented in the work of Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz (1998) and also by Tiago Alves Almeida, José María Gómez Hidalgo, and Akebo Yamakami (2011), our method provides considerable enhancements. These classical models are based on manually engineered attributes like word occurrence or term frequency-inverse document frequency (TF-IDF) values. However, these features do not fully capture the semantic or sequential characteristics of a language. In fact, this has been observed in studies such as those conducted by Rafiqul Islam, Rumana Islam, and Md. Monjurul Islam Rahman in 2020 and Enrico Blanzieri and Anton Bryl in 2008, that the conventional models face difficulty in context-sensitive classification. By contrast, our approach automatically learns representations directly from the data and dynamically emphasizes the relevant parts of the message through the attention mechanism. This leads to better generalization, improved accuracy, and greater practical applicability in real-world spam detection systems, where context awareness and adaptability are crucial.

4. Results

To compare the performance of the model suggested with an attention mechanism, we tested and trained the model on a dataset of pre-tagged email messages classified as spam and ham. The dataset was partitioned into 80% for training and 20% for testing. The training session was performed for 10 epochs utilizing a specified batch size of 32. The model performed as follows:

Measurement	**Value**
Accuracy	0.9980
Precision	0.9977
Recall	1.0000
F1-Score	0.9988

Table1: Training Results

The model worked very well on the training set, implying that it learnt the semantic and sequential patterns of unsolicited and legitimate messages well. The high precision implies that the model is capable of properly Categorizing spam textual data without Categorizing Many legitimate emails (ham) as spam. The high recall implies that few spam messages are missed.

Measurement	**Value**
Accuracy	0.9791
Precision	0.9785
Recall	0.9907

F1-Score	0.9846

Table2: Testing results

An interface that provides a user-to-input interface enabling the user to type in an email message at real-time was implemented. It takes the text entered, utilizes the same tokenization and padding procedures, and passes it on through the pre-trained model that calculates the chance of the email being spam. The feature has proved the capability of the model in real environments with real-email systems.

To contextualize the efficacy of our method, we compared the results of our deep learning model with classical machine learning algorithms from key literature:

Measurement	**Accuracy**	**Precision**
Naïve Bayes	0.9270	0.9010
SVM	0.9450	0.9200
Decision Tree	0.9370	0.9100
Proposed LSTM with Attention	0.9731	0.9785

Table 3: Comparison of Accuracy and Precision with Existing Methods

As exemplified, LSTM with attention gains evident performance advancement compared to the standard machine learning model. Handcrafted attributes like word frequency or term frequency-inverse document frequency (TF-IDF) are most common in standard models, lacking semantic subtleties and order of words. In contrast, the deep model automatically learns interpretable patterns and attention emphasizes contextual critical words.

5. Conclusion

In this paper, we introduced an LSTM network-based spam email detection model improved incorporating an attention mechanism. This design helps the model to efficiently learn both the sequential behaviour of text and the context-dependent significance of each word—enabling it to accurately identify spam messages from genuine ones with precision and reliability.

The model was then trained and validated on a marked-up collection of emails and performed very well, with training accuracy at 99.80% and test accuracy at 97.31%. The attention mechanism also increased the interpretability of the model by pointing out the most significant words that go into making decisions about classification, including "free," "win," and "urgent."

In contrast to conventional Machine learning models including Naïve Bayes and Support Vector Machines, which are based on manually designed features, our approach learns from raw data directly and exhibits better generalization to new inputs. Besides, a real-time prediction interface was also introduced, which illustrated the model's usability in actual applications.

In future research, this method can be extended by adding Bidirectional LSTM or Transformer-based models to further improve performance. Adding multilingual emails to the dataset and constantly changing spam patterns can also enhance robustness and flexibility.

6. References

[1] T. S. Guzella and W. M. Caminhas, "A Review of Machine Learning Approaches to Spam Filtering," Expert Systems with Applications, vol. 36, no. 7, pp. 10206–10222, 2009.

[2] T. A. Almeida, J. M. Gómez Hidalgo, and A. Yamakami, "Spam Filtering: How the Dimensionality Reduction Affects the Accuracy of Naïve Bayes Classifiers," Journal of Internet Services and Applications, vol. 1, no. 3, pp. 183–200, 2011.

[3] R. Islam, R. Islam, and M. M. Rahman, "A Comparative Study on Email Spam Detection Using Different Machine Learning Models," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 2, pp. 518–523, 2020.

[4] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian Approach to Filtering Junk E-Mail," in Proc. AAAI Workshop on Learning for Text Categorization, Madison, WI, USA, Jul. 1998, pp. 62–69.

[5] Blanzieri, E., & Bryl, A. (2008). A comprehensive review of machine learning techniques used in email spam filtering. Artificial Intelligence Review, 29(1), 63-92.

[6] AbdulNabi, I., & Yaseen, Q. (2021). Utilizing deep learning methods for spam email detection. ResearchGate.

[7] Ragesh, T. S., Tamizhselvi, A., & Patil, H. (2023). Enhancing email spam detection through a hybrid BiGRU-A-LSTM approach. ResearchGate.

[8] Ullah, S., & Rahman, S. E. (2020). Spam email classification using a combination of Bidirectional LSTM and Convolutional Neural Networks. ResearchGate.

[9] Tusher, E. H., Ismail, M. A., & Raffei, A. F. M. (2022). Email spam detection with LSTM networks enhanced by word embedding techniques. ResearchGate.

[10] SCIRP. (2023). A deep neural network approach for spam email classification with attention mechanisms. SCIRP.