IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

AI BASED SPAM SPOILER FOR E-MAIL SERVICES

[1]Ms ARUNA T N, [2]REESHMA R, [3]PRIYANKA R, [4]RAJAHA MUTHIAHA C, [5]THANUSU C

[1]Assistant Professor, [2]UG Student, [3]UG Student, [4]UG Student, [5]UG Student
Department of Computer Science Engineering
KGISL INSTITUTE OF TECHNOLOGY COIMBATORE INDIA

ABSTRACT : ***In recent times, cyber cover incidents have came down constantly. In utmost of these incidents, the bushwhacker used colorful types of spam emails as a driving force and successfully compromised her websites of government systems, well-known companies, politicians and social associations in numerous countries. increase. The discovery of spam emails from a large quantum of dispatch data is attracting attention. still, spam dispatch disguise ways are getting decreasingly sophisticated, and being discovery styles are ill- equipped to keep pace with decreasingly sophisticated fraud ways and the volume of emails. This design develops a new and effective approach called Spam Spoiler, which uses LSTM- grounded GRUs to classify large quantities of dispatch data into four distinct classes generally fraudulent, draining, and suspicious emails. suggested to do The new process includes two crucial phases. A sample expansion phase and a test phase with enough samples. An LSTM- grounded GRU, this design efficiently retrieves meaningful information from emails that can be used as substantiation for forensic analysis. Experimental results show that Spam Spoiler outperforms being ML algorithms, achieving 98 bracket delicacy using his LSTM novel fashion with iterative grade unit. Dispatch content analysis covers different types of motifs. Spam Spoiler effectively outperforms being styles, leaving the bracket process robust and dependable.***

Keyword -AI,E-Mail, LSTM, GRU, Spam Detection

CHAPTER 1
INTRODUCTION

1.1. General

E-mail is short for e-mail. This is a method of sending messages from one computer to another over the Internet. It is primarily used in the areas of business, education, technical communication, and document interaction. Enables communication without disturbing people around the world. In 1971, Ray His Tomlinson sent himself a test email containing the text. Email His messages are sent via an email server. Use multiple protocols within the TCP/IP suite. For example, SMTP is a protocol, short for Simple Mail Transfer Protocol, and is used to send messages, while other protocols such as IMAP and POP are used to retrieve messages from mail servers. To log into your email account, simply enter a valid email address, password, and the email server you use to send and receive messages. Most webmail servers set up your email account automatically, but you only need to enter your email address and password. However, if you use an email client such as Microsoft Outlook or Apple Mail, you may need to manually configure each account. Additionally, entering your e-mail address and password may also require you to enter your incoming and outgoing mail servers and the correct port number for each.

1.2. Scope of the project

1. Existing approaches to email classification result in irrelevant emails and loss of valuable information.
2. The scope of this project is to design a new efficient approach called Email Sink API for classifying emails into four different classes: normal, fraudulent, threatening and suspicious emails using an LSTM-based GRU.
3. Email Sink API focuses on optimizing LSTM-based GRU parameters for best performance.
4. LSTM-based GRU efficiently retrieves meaningful information from emails that can be used as evidence for forensic analysis.
5. Email content analysis helps identify spoofing because it is more efficient to analyze the headers of specific emails than all emails.

1.3. Objectives of the project

1. Investigating crimes involving electronic mail (email) requires analysis of both the email header and body.
2. So the semantics of the communication help identify possible sources of evidence.
3. Choosing the best model for email forensics tools.

4. A proposal for a new efficient approach called Email Sink API that uses a Long Short-Term Memory (LSTM) based Gated Recurrent Neural Network (GRU) for multi-class email classification.
5. Based on deep learning-based architecture, it detects all malicious or unwanted emails received on the email server side.
6. Simultaneously model emails at the email header, email body, character level, and word level.
7. To identify whether an email is a cybercrime email.

CHAPTER 2
LITERATURE SURVEY

[1] The goal is to design a new SDS model that achieves promising scores in terms of classification accuracy, detection rate, false alarm rate, and global convergence, and shows strong spam detection power using six different improved Grasshopper optimization variants. Algorithm (EGOA) for training ANNs. MLPs are powerful classifiers that have proven to be very capable in solving SDS challenges. Such a tool identifies typical characteristics of system users, email context, and statistically significant deviations from established user behavior.
Methodology – SSD model based on ANN.

[2] The system goes through two phases: training and testing. The training phase has four modules: data preparation, feature selection, feature reduction, and classification. The testing phase consists of data preparation and classification modules. A decision tree uses a tree-like model to represent several possible decision paths and their possible outputs. In a decision tree, each node represents an attribute, each branch represents a decision, and each leaf represents an outcome (class or decision). Decision trees can be used to predict the class of an unknown query event by creating a model trained on labeled data. Each training example should be characterized by several descriptive properties or characteristics. Properties can have either nominal values or constant values.
Methodology – Spam Detection (SD)system.

CHAPTER 3
ALGORITHMS

3.1. LSTM [Long Short- Term Memory]

LSTMSs are a special kind of RNN designed to learn long-term dependencies. They're composed of a cell, an input gate, an affair gate, and a forget gate. The cell is responsible for flashing back values over arbitrary time intervals, while the three gates are used to regulate the information to be kept or discarded at circle operation before passing on the long-term and short-term information to the coming cell. LSTMs were developed to deal with the exploding and evaporating grade problem when training traditional RNN. There is an aggregate of three gates that LSTM uses the Input Gate, Forget Gate, and Affair Gate.

3.1.1. Input Gate

The input port decides what information is stored in long-term memory. It only works with the data from the current input and the short-term memory from the previous step. With this port, it filters data from variables that are not useful.

3.1.2. Forget Gate

Forget decides what information to keep or discard from long-term memory by multiplying the incoming long-term memory with the forget vector generated from the current input and the incoming short-term memory.

3.1.3. Output Gate

The output gate uses the current input, the previous short-term memory, and the newly calculated long-term memory to create a new short-term memory that is transferred to the cell at the next time step. The result of the current time step can also be drawn from this hidden space.

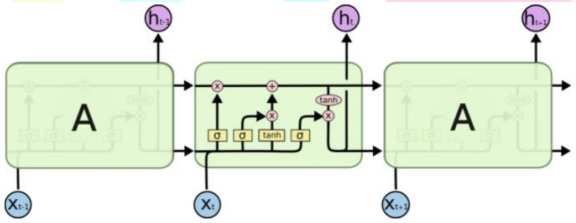

Fig.3.1 input, forget, output gates

3.2. GRU [Gated Recurrent Unit]

The Gated intermittent Unit (GRU) is a gating medium in intermittent neural networks (RNN) analogous to a long short- term memory (LSTM) unit but without an affair gate. It solves the evaporating grade problem by using an update gate and a reset gate, which can be trained to keep information from the history or remove in applicable information. GRUs have better performance than LSTM when dealing with lower datasets. The workflow of the GRU is the same as the RNN but the difference is in the operation and gates associated with each GRU unit.

3.2.1. Update Gate

The update gate is responsible for determining how much previous information should be passed to the next state. This is very powerful because the model can decide to copy all the information from the past, eliminating the risk of vanishing gradients.

3.2.2. Reset Gate

The reset gate stores relevant information from the past time step into new memory content, multiplies the input vector and hidden state with their weights, and calculates element-wise multiplication.

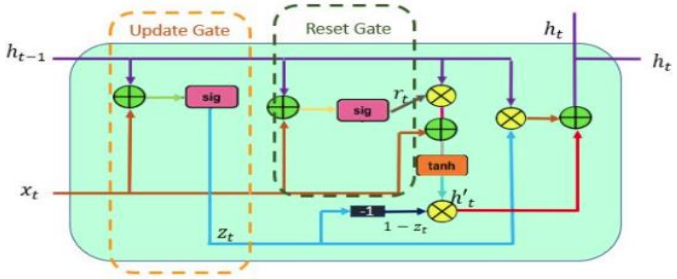

Fig. 3.2 update, reset gate

CHAPTER 4
PROJECT DESCRIPTION

Email messages contain two main components: the header and the body. The header contains content-related information, while the body contains variable contents. Pre-processing is done to remove URLs, HTML, CSS, JavaScript code, and special symbols. Deep machine learning is used for text-based classification. The tokenization layer maintains a dictionary that maps a word to an index, while the embedding layer maintains a lookup table that maps the index/token to a vector. Two LSTM networks feed in the input sequence, and the merge of these two networks serves as input to the next layer. The Glove word embedding from Stanford NLP Group was used to apply the Word Cloud embeddings.

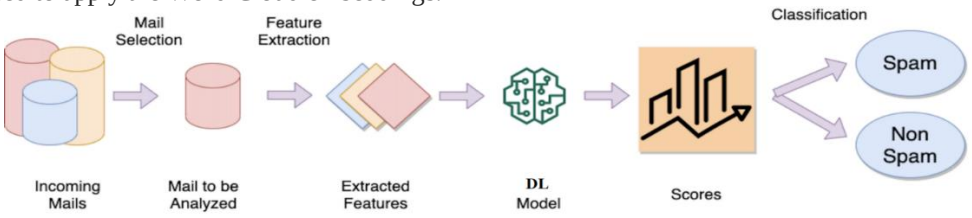

Fig. describe project flow

4.1. Dataset Description

The dataset used in this project is an amalgamation of four different datasets: Normal e-mails from Enron Corpora, Fraudulent e-mails from Phished e-mails corpora, Harassment messages from Hate Speech, and Offensive e-mails. Suspicious emails data from email sources and Twitter source are added to make multiclass E-mail classification possible.

Table 4.1 Composition of Dataset

No.	Class Name	No. of E-Mails
1	Fraudulent	9001
2	Harassment	9138
3	Suspicious	5287
4	Normal	9001

4.2. Modules Description

4.2.1. E-Mail Sink AI Web App

The E-mail Forensics Predictor service is an online platform for people to use emails free from spam. It integrates with Trainer and Tester Modules developed with Python and Flask Framework.

4.2.1.1. User Account Management

1. Admin
2. Email user

4.2.2. E-Mail Classification – Training Phase

4.2.1.1. Dataset Annotation

The E- correspondence is divided into word situations of the E- correspondence body, and the embedding estate is used to train and gain the sequence of vectors.

4.2.2.2. Data Pre-processing

The data pre-processing phase consists of natural language-based steps that standardize the text and prepare it for analysis.

1. **Tokenization**

Breaking up the original text into element pieces is the tokenization step in natural language processing. There are predefined rules for tokenization of the documents into words. Tokenization is performed in Python using the SpaCy library.

2. **Normalization**

These are the way took for rephrasing textbook from mortal language (English) to machine- readable format for farther processing. The process starts with the most important idea is to change all rudiments to lower or upper case. expanding bowdlerization banning figures or changing those to words removing white spaces removing punctuation, curve marks, and other circumflexes Removing stop words, meager terms, and particular words is essential for effective jotting. textbook canonicalization

3. **Stop Words Removal**

Stop words, such as a" and the", are not relevant to the context of the E-mail and create noise in the text data. We used the NLTK Python library to remove stop words from the text.

4. **Punctuation Removal**

Punctuation includes (e.g., full stop (.), comma (,), brackets) to separate sentences and clarify meaning. For punctuation removal, we utilize the ``NLTK" library.

6.2.2.4. Feature Extraction

The TF- IDF system is used to convert an developed list of words into figures. A popular and straightforward system of point birth with textbook data is the bag- of- words model. It involves two effects a vocabulary of given words and a measure of the presence of given words. It excerpts features on the base of Equations, where if it represents term frequency and ds represents document frequency

$$TFIDF = tf * (\frac{1}{df})$$
$$TFIDF = tf * Inverse(df)$$
$$TFIDF(t, d, D) = TF(t, d).IDF(t, D)$$
$$TFIDF(t, d) = log\frac{N}{|d \epsilon Dt \epsilon D|}$$

Eq. 1-5

Word2vec is a neural network- grounded algorithm used for point birth in DL with the environment of words. Equation 5 shows how it manages the word- environment with probability measures

$$P(D = 1 \mid w, c_{1:k}) = \frac{1}{1 + e^{-(w \cdot c_1 + w \cdot c_2 + \ldots + w \cdot c_k)}}$$

Eq.5

Multi-word environment is a variant of word2vec. The variable- length environment is also controlled by the given below mathematics.

$$P(D = 1 \mid w, c) = \frac{1}{1 + e^{-s(w,c)}}$$

Eq.6

Fig. 4.3.2.4 Word Embedding Layer

4.2.2.5. Word Embedding Layer

Word embedding is the representation of words into real numbers. It uses a word mapping dictionary to convert the terms (words) to a real value vector. There are two main problems with machine learning feature engineering techniques, one being the sparse vectors and the other not taking into account the meaning of words. In the experimental setup, the word embedding layer contains information about the sequence length of E-mails. The embedding dimensions used in SeFACED are 800, the vocabulary size is 70, 000, and the input length is 600. The embedding layer takes three arguments such as input dimensions, output dimensions, and input length. The embedding layer output will be used for the LSTM and GRU layers in adjacent layers.

4.2.2.6. LSTM based GRU Classification Model

LSTMs typically include email classifications such as harassment, suspicious, and fraud. Since both LSTM and GRU are based on gated network architecture, we combined GRU and LSTM to take advantage of both gated architectures. The hierarchical structure of DL models helps them learn without interfering with the implementation of ML models. Some libraries provide detailed learning implementation structures. We split the data into three sets: training, validation, and testing, with a ratio of 65V vs. 10V vs. 25. We used word embedding techniques to extract features from email text data. Encode the target values into four different classes using the one-hot encoding technique. To fully classify emails, we pass all preprocessed data through a new architecture of LSTM layer variants. We use LSTM layers with different GRU and Convo1D layer variants to transform input text data into efficient email classification system.

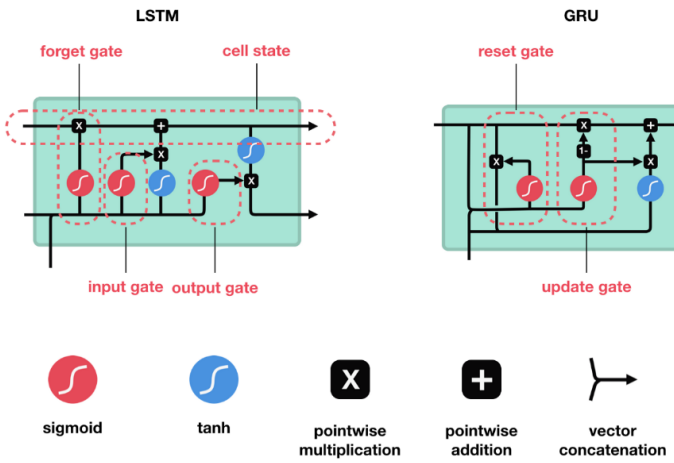

Fig. LSTM based GRU Classification Model

Textual data needs special attention when feature extraction comes in the proposed methodology. Different feature extraction methods need to be implemented when solving the Natural language processing problem using DL. The main point is to convert the textual data into real-valued vectors. There are multiple ways to generate the embedding vector from the textual data, but famous methods are GLOVE and Word2Vec techniques. The dimensions of the embedding vector are essential to get all the features extracted from the data. For the classification problem of NLP, the target values need to be encoded using the one-hot encoding method. After getting the vectors from the words, the similarity between the words is measured using the similarity measure between the corresponding vectors using equation 7

$$\text{sim}_{\cos}(u, v) = \frac{u \cdot v}{\|u\|_2 \|v\|_2} = \frac{\sum_i u_{[i]} \cdot v_{[i]}}{\sqrt{\sum_i (u_{[i]})^2} \sqrt{\sum_i (v_{[i]})^2}}$$

Eq.7

There are many other ways to measure the similarity between the word vectors, one of them is Jaccard similarity, which defines Equation 8.

$$\text{sim}_{\text{Jacard}}(u, v) = \frac{\sum_i \min(u_{[i]}, v_{[i]})}{\sum_i \max(u_{[i]}, v_{[i]})}$$

Eq.8

DL for NLP uses dense vector representation to reduce memory requirements and categorical data encoding, but most literature is based on one-hot encoding techniques. After feature extraction, language modelling follows.

4.3.3. E-Mail Forensic Analyzer – Testing Phase

4.3.3.1. E-Mail Forensic Predictor

This module predicts whether it is spam or not spam and classify the type. Block the mail and send alert message to the user if the mail is spam.

Fid 4.3.3.1. Decision Making

CHAPTER 5
EVALUATION METRICS

4.1. Confusion Matrix

Spam detection can be evaluated using various performance metrics. A confusion matrix is used to visualize the email detection of the models. Several metrics such as precision, accuracy, recall, and f-score are used to evaluate the performance of classifiers. These measurements are calculated using a four-period confusion matrix.

Confusion matrix can be defined as below:

1. True positive (TP): are the positive values correctly classified as positive.
2. True Negative (TN): are the negative values correctly classified as negative.
3. False Positive (FP): are the negative values incorrectly classified as positive.
4. False Negative (FN): are the positive values incorrectly classified as negative.

For the performance evaluation of our proposed model, we use the following metrics.

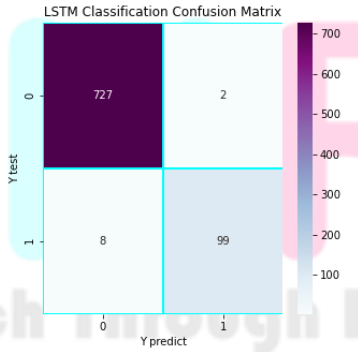

Fig.4.1 LSTM classification confusion Matrix

4.2. Accuracy

The Accuracy of a discovery medium can be calculated using Equation.

Accuracy = TP+TN/TP+TN+FP+FN

training score :0.991249719542293

testing score :0.979372197309417

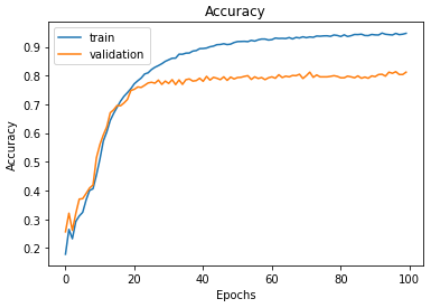

Fig.4.2 Precision

Is the bit of the predicted correctly classified operations to the total of all operations that are correctly real positive? It can be calculated using Equation 13.

Precision=TP/TP+FP

8.1.4. Recall

The recall is a bit of the predicted correctly classified operations to the total number of operations classified correctly or erroneously. Recall can be calculated using Equation 14.

Recall=TP/TP+FN

8.1.5.F-Score

F- score is the harmonious mean of perfection and recall. The model's ability to distinguish between different objects is symbolized.. f- score of a discovery model can be reckoned using Equation 15.

F1-Score=2*(Precision*Recall)/(Precision + Recall)

CHAPTER 5
RESULTS AND DISCUSSION

The results and comparisons of different classifiers after training and testing on data. We collected 5000 emails from the online resource 'Kaggle' and translated them into Urdu using the Google trans-Python library using the Google Translate Ajax API. 4,000 emails were used to train various ML and DL models. 1,000 emails were used for testing to quantify accuracy and metrics. Accuracy, precision, recall, and f-measures were assessed as described in section 5 on rating scales. These are SVM and Naive Bayes, LSTM, and GRU are used to measure ROC AUC and model loss values. Finally, below is a comparison of the models using different graphics. The results in Table 4 show that the deep learning algorithm (LSTM) is a powerful method to detect Urdu spam emails with a high accuracy of 98.4%.

Table : Accuracy of different models

Models	Accurancy(%)
LSTM	98.4
CNN	96.2
Naïve Bayes	98.0
SVM	97.5

In the above table, we compared the accuracy of four different ML and DL models. We find that the DL model (LSTM) is the most accurate of all models, but takes longer to train. ML models such as SVM and Naive Bayes have lower accuracy percentages than LSTM/CNN, which are also DL models and have the lowest accuracy percentages.

CHAPTER 6
CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

Post-incident proactive monitoring and analysis of email data is important for organizations as cybercrime and incidents due to vulnerabilities are on the rise. Cybercrimes such as hacking, spoofing, phishing, email bombing, whaling and spam are carried out through email. Existing email classification approaches result in irrelevant emails and loss of valuable information. With these limitations in mind, we developed a new efficient approach called E-Mail Sink AI, which uses an LSTM-based GRU that handles not only short sequences, but also handles, to treat email as normal, fraudulent, threatening, and We have categorized them into four different classes of suspicious emails. 1000C character long dependencies. We evaluated his proposed Email Sink AI model using metrics such as Precision, Recall, Accuracy, and F-Score. Experimental results show that E-Mail Sink AI outperforms existing ML algorithms and achieves 95% classification accuracy using a novel method of LSTM with iterative gradient unit.

6.2 Future Scope

1. At this time, we typically consider email classes such as harassing, fraudulent, and suspicious.
2. However, given the sheer volume of email data, many other classes could be added to this task.

CHAPTER 7
REFERENCE

[1] S. Sinha, I. Ghosh, and S. C. Satapathy, ``A study for ANN model for spam classification," in Intelligent Data Engineering and Analytics. Singapore: Springer, 2021, pp. 331-343.

[2] Q. Li, M. Cheng, J. Wang, and B. Sun, ``LSTM based phishing detection for big email data," IEEE Trans. Big Data, early access, Mar. 12, 2020.

[3] T. Gangavarapu, C. D. Jaidhar, and B. Chanduka, ``Applicability of machine learning in spam and phishing email filtering: Review andapproaches," Artif. Intell. Rev., vol. 53, no. 7, pp. 5019-5081, Oct. 2020.

[4] E. Bauer. 15 Outrageous Email Spam Statistics That Still Ring True in 2018, RSS. Accessed: Oct. 10, 2020.
[5] A.Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, ``A comprehensive survey for intelligent spam email detection," IEEE Access, vol. 7.

International Research Journal
IJNRD
Research Through Innovation