Journal of Positive School Psychology http://journalppw.com
2022, Vol.6, No.4, 5425-5432

E-mail Sink AI: Deep Learning model for multiclass E-mail Classification for Forensic Analysis using Design Thinking.

Dr.P.Sumathi [1] , R. Elavarasan*, M.Sadhakrishnan* , K.Harishanand*
Head Of the Department[1] , Student*,
Department of Information Technology,
SNS College Of Engineering, Coimbatore, India
Hod.it@snsce.ac.in ,elaelavarasan762@gmail.com , sadhakrishnanm@gmail.com , harishanand1820@gmail.com

Abstract

E-mail is an essential application for carrying out transactions and efficiency in business processes to improve productivity. E-mail is frequently used as a vital medium of communication and is also being used by cybercriminals to commit crimes. Cybercrimes like hacking, spoofing, phishing, E-mail bombing, whaling, and spamming are being performed through E-mails. Hence, there is a need for proactive data analysis to prevent cyber-attacks and crimes. Keeping in sight these limitations, this project proposed to design a novel efficient approach named E-mailSinkAI for E-mail classification into four different classes: Normal, Fraudulent, Threatening, and Suspicious E-mails by using LSTM based GRU. The LSTM based GRU efficiently captures meaningful information from E-mails that can be used for forensic analysis as evidence. E-mailSinkAI effectively outperforms existing methods while keeping the classification process robust and reliable.

Introduction

1.1. Overview

Email stands for Electronic Mail. It is a method to sends messages from one computer to another computer through the internet. It is mostly used in business, education, technical communication, document interactions. It allows communicating with people all over the world without bothering them. In 1971, a test email sent Ray Tomlinson to himself containing text.

Figure 1.1. E-Mail

Email messages are conveyed through email servers; it uses multiple protocols within the TCP/IP suite. For example, SMTP is a protocol, stands for simple mail transfer protocol and used to send messages whereas other protocols IMAP or POP are used to retrieve messages from a mail server. If you want to login to your mail account, you just need to enter a valid email address, password, and the mail servers used to send and receive messages.

Email messages include three components, which are as follows:

- Message envelope: It depicts the email's electronic format.
- Message header: It contains email subject line and sender/recipient information.
- Message body: It comprises images, text, and other file attachments.

1.1.1. Types of Email

1. Newsletters

It is a type of email sent by an individual or company to the subscriber. It contains an advertisement, product promotion, updates regarding the organization, and marketing content. It might be upcoming events, seminars, webinars from the organization.

2. Onboarding emails

It is an email a user receives right after subscription. These emails are sent to buyers to familiarize and tell them about to use a product. It also contains details about the journey in the new organization.

3. Transactional

These types of emails might contain invoices for recent transactions, details about transactions. If transactions failed then details about when the amount will be reverted. We can say that transaction emails are confirmation of purchase.

4. Plain-Text Emails

These types of emails contain just simple text similar to other text message services. It does not include images, videos, documents, graphics, or any attachments. Plain-text emails are also used to send casual chatting like other text message services.

1.2. Problems Identified

Many people rely on the Internet for many of their professional, social and personal activities. But there are also people who attempt to damage our Internet-connected computers, violate our privacy and render inoperable Internet services.

Email is a universal service used by over a billion people worldwide. As one of the most popular services, email has become a major vulnerability to users and organizations. The statistics are astounding. Email remains the number one threat vector for data breaches, the point of entry for ninety-four percent of breaches. There is an attack every 39 seconds. Over 30% of phishing messages get opened, and 12% of users click on malicious links. As cybercrime becomes more advanced and bypasses the legacy controls put in place to defend against it, security must become more advanced too.

Figure 1.2. E-Mail Attacks

Below are some of the most common types of Attacks:

1. **Phishing:**

Phishing is a form of fraud. Cyber criminals use email, instant messaging, or other social media to try to gather information such as login credentials by masquerading as a reputable person. Phishing occurs when a malicious party sends a fraudulent email disguised as being from an authorized, trusted source. The message intent is to trick the recipient into installing malware on his or her device or into sharing personal or financial information.

2. **Vishing:**

Vishing is phishing using voice communication technology. Criminals can spoof calls from authorized sources using voice-over IP technology. Victims may also receive a recorded message that appears authorized. Criminals want to obtain credit card numbers or other information to steal the victim's identity. Vishing takes advantage of the fact that people trust the telephone network.

3. **Smishing:**

Smishing is phishing using text messaging on mobile phones. Criminals impersonate a legitimate source in an attempt to gain the trust of the victim. For example, a smishing attack might send the victim a website link. When the victim visits the website, malware is installed on the mobile phone.

4. **Whaling:**

Whaling is a phishing attack that targets high profile targets within an organization such as senior executives. Additional targets include politicians or celebrities.

5. **Pharming:**

Pharming is the impersonation of an authorized website in an effort to deceive users into entering their credentials. Pharming misdirects users to a fake website that appears to be official. Victims then enter their personal information thinking that they are connected to a legitimate site.

6. **Spyware:**

Spyware is software that enables a criminal to obtain information about a user's computer activities. Spyware often includes activity trackers, keystroke collection, and data capture. In an attempt to overcome security measures, spyware often modifies security settings. Spyware often bundles itself with legitimate software or with Trojan horses. Many shareware websites are full of spyware.

1.3. Objective of the Project

To investigate crimes involving Electronic Mail (e-mail), analysis of both the header and the email body is required since the semantics of communication helps to identify the source of potential evidence.

To select the best model for E-mail forensic tools.

To propose a novel efficient approach named E-MailSinkAPI that uses Long Short-Term Memory (LSTM) based Gated Recurrent Neural Network (GRU) for multiclass email classification.

To detect any harmful or unfavourable e-mails received at the e-mail server end based on the deep learning-based architecture.

To model emails at the email header, the email body, the character level, and the word level simultaneously.

To distinguish whether the email is a cybercrime email.

Existing System

1. **Content-Based Filtering Technique**

Algorithms analyse words, the occurrence of words, and the distribution of words and phrases inside the content of e-mails and segregate them into spam non-spam categories.

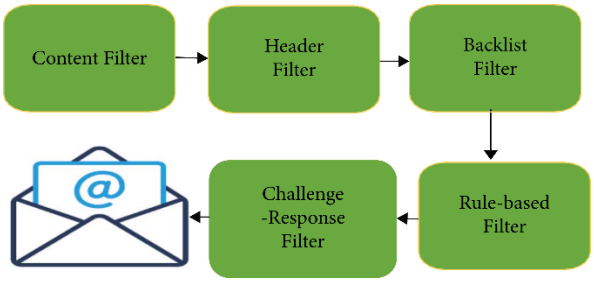


Figure 3.1. Content based Email Filtering

2. Case Base Spam Filtering Method

Algorithms trained on well-annotated spam/non-spam marked emails try to classify the incoming mails into two categories.

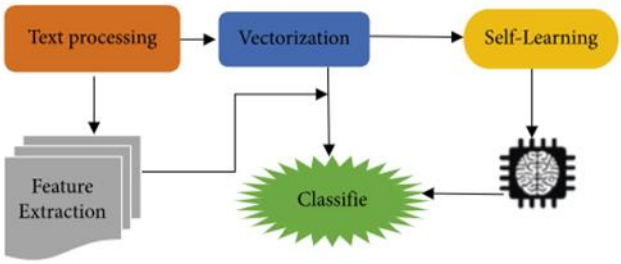


Figure 3.2. Case based Spam Filtering

3. **Heuristic or Rule-Based Spam Filtering Technique**

Algorithms use pre-defined rules in the form of a regular expression to give a score to the messages present in the e-mails. Based on the scores generated, they segregate emails into spam non-spam categories.

4. The Previous Likeness Based Spam Filtering Technique

Algorithms extract the incoming mails' features and create a multi-dimensional space vector and draw points for every new instance. Based on the KNN algorithm, these new points get assigned to the closest class of spam and non-spam.

5. Adaptive Spam Filtering Technique

Algorithms classify the incoming mails in various groups and, based on the comparison

scores of every group with the defined set of groups, spam and non-spam emails got segregated.

2.1 Proposed System

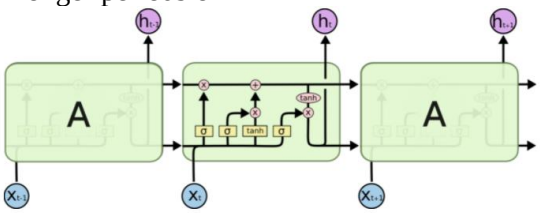
The proposed approach comprises data collection, pre-processing, feature extraction, parameter tuning, and classification using the LSTM-GRU model. In this project, E-mail datasets are divided into normal, harassing, suspicious, and fraudulent classes. The E-mail is divided into word levels of the E-mail body, and the embedding layer is applied to train and obtain the sequence of vectors.

2.1.1 LSTM and GRU

In Deep learning, Long-Term Short-Term Memory Networks and Gated Recurrent Units, LSTM and GRUs for short.

- **LSTM – Long Short-Term Memory**

LSTMS are a special kind of RNN which is capable of learning long-term dependencies. LSTMs are designed to dodge long-term dependency problem as they are capable of remembering information for longer periods of time. Long short-term memory (LSTM) units (or blocks) are a building unit for layers of a recurrent neural network (RNN). A RNN composed of LSTM units is often called an LSTM network. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell is responsible for "remembering" values over arbitrary time intervals; hence the word "memory" in LSTM. Each of the three gates can be thought of as a "conventional" artificial neuron, as in a multi-layer (or feedforward) neural network: that is, they compute an activation (using an activation function) of a weighted sum. Intuitively, they can be thought as regulators of the flow of values that goes through the connections of the LSTM; hence the denotation "gate". There are connections between these gates and the cell. LSTMs were developed to deal with the exploding and vanishing gradient problem when training traditional RNNs.



The popularity of LSTM is due to the Getting mechanism involved with each LSTM cell. In a normal RNN cell, the input at the time stamp and hidden state from the previous time step is passed through the activation layer to obtain a new state. Whereas in LSTM the process is slightly complex, as you can see in the above architecture at each time it takes input from three different states like the current input state, the short-term memory from the previous cell and lastly the long-term memory. There are a total of three gates that LSTM uses as Input Gate, Forget Gate, and Output Gate.

Input Gate

The input gate decides what information will be stored in long term memory. It only works with the information from the current input and short-term memory from the previous step. At this gate, it filters out the information from variables that are not useful.

Forget Gate

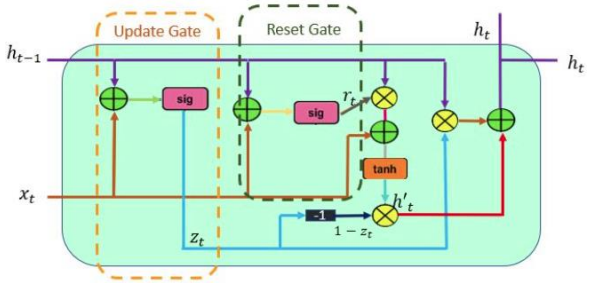
The forget decides which information from long term memory be kept or discarded and this is done by multiplying the incoming long-term memory by a forget vector generated by the current input and incoming short memory.

Output Gate

The output gate will take the current input, the previous short-term memory and newly computed long-term memory to produce new short-term memory which will be passed on to the cell in the next time step. The output of the current time step can also be drawn from this hidden state.

- **GRU – Gated Recurrent Unit**

Gated recurrent unit (GRU) was introduced by Cho, et al. in 2014 to solve the vanishing gradient problem faced by standard recurrent neural networks (RNN). GRU shares many properties of long short-term memory (LSTM). Both algorithms use a gating mechanism to control the memorization process. A gated recurrent unit (GRU) is a gating mechanism in recurrent neural networks (RNN) similar to a long short-term memory (LSTM) unit but without an output gate. GRU's try to solve the vanishing gradient problem that can come with standard recurrent neural networks.



The workflow of the Gated Recurrent Unit, in short GRU, is the same as the RNN but the difference is in the operation and gates associated with each GRU unit. To solve the problem faced by standard RNN, GRU incorporates the two gate operating mechanisms called Update gate and Reset gate.

Reset gate

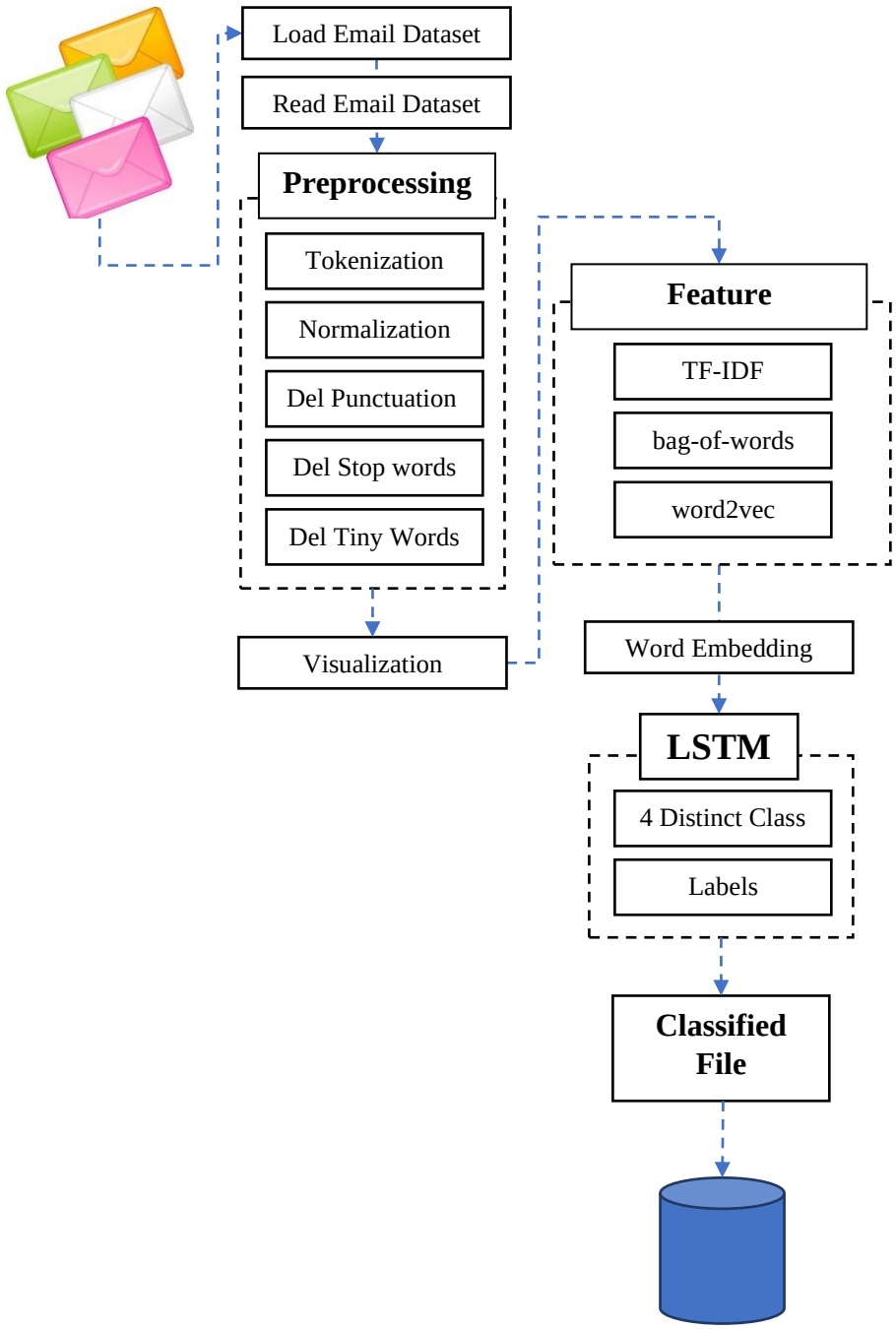
The reset gate is used from the model to decide how much of the past information is needed to neglect; in short, it decides whether the previous cell state is important or not.

First, the reset gate comes into action it stores relevant information from the past time step into new memory content. Then it multiplies the input vector and hidden state with their weights. Next, it calculates element-wise multiplication between the reset gate and previously hidden state multiple. After summing up the above steps the non-linear activation function is applied and the next sequence is generated.

Update gate

The update gate is responsible for determining the amount of previous information that needs to pass along the next state. This is really powerful because the model can decide to copy all the information from the past and eliminate the risk of vanishing gradient.

System Architecture – Training Phase



System Architecture
Phase
Testing Phase

3. Modules Description

3.1 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 component pieces is the tokenization step in natural language processing. There are predefined rules for tokenization of the documents into words. The tokenization step is performed in Python by using the SpaCy library.

2) STOP WORDS REMOVAL

Words like ``a'' and ``the'' that appear so frequently are not relevant to the context of the E-mail and create noise in the text data. These words are called stop words, and they can be filtered from the text to be processed. We utilized the``NLTK'' Python library to remove stop words from the text.

3) PUNCTUATION REMOVAL

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

3.2 Feature Extraction

After eliminating irrelevant information, the elaborated list of words is converted

The TF-IDF method is applied to accomplish this task. Term Frequency is several occurrences of a word in a document, and IDF is the ratio of a total number of documents and the number of documents containing the term. A popular and straightforward method of feature extraction with text data is called the bag-of-words model of text. A bag-of-words model, or BoW for short, is a way of extracting features from the text for use in modelling, such as machine learning algorithms. A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things (1) A vocabulary of known words, (2) A measure of the presence of known words. We extract features on the basis of Equations Here tf represents term frequency and df 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

Feature extraction in DL with the context of words is also essential. The technique used for this purpose is word2vec neural network-based algorithm. Equation 5 given below shows how word2vec manages the word-context with the help of probability measures. The D represents the pair-wise illustration of a set of words, and (w; c) is the word-context pair drawn from the large set D.

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

Eq.5

The multi-word context is also a variant of word2vec, as shown in Equation 6. The variable-length context is also controlled by the given below mathematics.

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

Eq.6

4. Conclusion

Several measurements are used for performance evaluation of classifiers like accuracy, precision, recall, and f-score. These measurements are computed by a confusion matrix, which is composed of four terms.

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

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

A. ACCURACY

Is the fraction of the total number of applications correctly classified? The Accuracy of a detection mechanism can be calculated using Equation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

B. PRECISION

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

$$Precision = \frac{TP}{TP + FP}$$

C. RECALL

The recall is a fraction of the predicted correctly classified applications to the total number of applications classified correctly or incorrectly. Recall can be calculated using Equation 14.

$$Recall = \frac{TP}{TP + FN}$$

D. F-SCORE

F-score is the harmonic mean of precision and recall. It symbolizes the capability of the model for making fine distinctions. f-score of a detection model can be computed using Equation 15.

$$F - score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

Reference

[1] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, ``A comprehensive survey of AI-enabled phishing attacks detection techniques,'' Telecommun. Syst., pp. 116, Oct. 2020.

[2] C. Iwendi, Z. Jalil, A. R. Javed, T. Reddy, R. Kaluri, G. Srivastava, and O. Jo, ``KeySplitWatermark: Zero watermarking algorithm for software protection against cyber-attacks," IEEE Access, vol. 8, pp. 7265072660, 2020.

[3] A. Rehman, S. U. Rehman, M. Khan, M. Alazab, and T. Reddy, ``CANintelliIDS: Detecting in-vehicle intrusion attacks on a controller area network using CNN and attention-based GRU," IEEE Trans. Netw. Sci. Eng., early access, Feb. 19, 2021, doi: 10.1109/TNSE.2021.3059881.

[4] T.R.Lekhaa, "Hacking The Information Using Phising" International Journal of Research in Computer Science 2014.

[5] S. U. Rehman, M. Khaliq, S. I. Imtiaz, A. Rasool, M. Shaq, A. R. Javed, Z. Jalil, and A. K. Bashir, ``DIDDOS: An approach for detection and identication of distributed denial of service (DDoS) cyberattacks using gated recurrent units (GRU)," Future Gener. Comput. Syst., vol. 118, pp. 453466, May 2021.

[6] S. I. Imtiaz, S. U. Rehman, A. R. Javed, Z. Jalil, X. Liu, andW. S. Alnumay, ``DeepAMD: Detection and identification of Android malware using high efficient deep artificial neural network," Future Gener. Comput. Syst., vol. 115, pp. 844856, Feb. 2021.