JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

E-MAILSINKAPI: DEEP LEARNING MULTICLASS E-MAIL CLASSIFICATION MODEL FOR FORENSIC ANALYSIS

***Mrs. A Sajitha Begam*[1], Abinandhan P[2], Mummurthi G[3], Naveen R[4]**

Ass.Professor[1], UG Student[234]

Department of Information Technology[1],

Adhiyamaan College of Engineering, Hosur, Tamil Nadu, India

Abstract: E-mail is a crucial tool for completing transactions and increases the performance of organizational techniques to increase productivity. Hacking, spoofing, phishing, E-mail bombing, whaling, and spamming are all e-mail-related cybercrimes. As a result, to prevent cyber assaults and crimes proactive records evaluation is essential. Because communication semantics aid in identifying the source of potential evidence, it's necessary to look into both the email header and the email body while looking into crimes committed by email. Investigators are now faced with the arduous task of extracting significant semantic information from large quantities of emails, which has caused the investigation to be postponed. Existing email classification methods result in the erasure of critical information and/or the transmission of unnecessary emails. Given these constraints, the research suggests that E-mailSinkAPI, a revolutionary efficient approach for categorizing e-mails into four distinct classes: regular, fraudulent, threatening, and suspicious, be constructed using Long Short-Term Unit (LSTM) based Grated Recurrent Unit (GRU). The GRU employs LSTM to extract critical data from emails that may be used as proof in forensic investigations E-mailSinkAPI outperforms previous technology even by retaining a constant and dependable categorization process.

Keyword: E-mailSinkAPI, GRU, LSTM, Classification

I. INTRODUCTION

The term "email" is used to describe electronic mail it is a process of transmitting data via the internet from one pc to another pc the most popular applications are in business education technical communication and document interaction. It enables individuals from throughout the world to converse without being harassed ray Tomlinson received a text-based test email in 1971.

Figure 1: E-Mail

Email data are sent using a variety of TCP/IP protocols by email servers for example the SMTP is for sending as well as receiving messages whereas IMAP and pop are for extracting data from an electronic mail server. To get entry to your email account, all you want is a demonstrated username, a password, and the mail servers to be used to ship and acquire messages. The objective of the systems is to explore crimes involving electronic mail (e-mail). Both the header and the email body must be scrutinized since communication semantics aid in recognizing the source of potential evidence.

1.1 Problems Identified

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

Figure 2: E-Mail Attacks

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.

II. RELATED WORKS

Sanaa A. A. Ghaleb**, (**2021) **"Enhance Grasshopper Optimization Algorithm for Spam Detection System Training Neural Networks"** [1] The data analysis presents a new Spam Detection System (SDS) framework for advanced spam email detection that is based on a set of six extended Grasshopper Optimization Algorithm (EGOA) variants that have been studied and merged with a Multilayer Perceptron (MLP). Where a suitable feature selection approach was not included, the system solely evaluated the models by considering all of the features of the spam detection datasets. As a result, in the future, the system will focus on constructing an effective SDS based on two goals: the first is to reduce the number of selected features, and the second is to combine the best-proposed models to create an ensemble-based spam detection classifier. Ivana Čavor**, (**2021) **"Classifying Emails Using a Decision Tree Model"** [2] The concept is to keep the maximum essential functions whilst reducing the quantity of processing essential after function choice the id3 set of rules is used to generate a selection tree that categorizes emails as unsolicited mail or ham. The recommended method is categorized: as performance, clarity, and memory. The suggested device's overall performance is classified in terms of the dataset and characteristic size. Where a suitable feature selection approach was not included, the system solely evaluated the models by considering all of the features of the spam detection datasets. Luis Felipe Gutiérrez, (2021) **"Phishing Detection Using Email Embeddings"** [3] The created group containing scamming x and legal maiemailsth identical symptoms to investigate if email embeddings, or vectorizations, catch or ignore these cues. The high overall classification results indicate that the semantic vector space in which the document vectors are stored is suitable for this classification task. Furthermore, the semantics of the content of the emails are ideally suited for class segmentation. The high overall classification results indicate that the semantic vector space in which the document vectors are stored is suitable for this classification task. Furthermore, the semantics of the content of the emails are ideally suited for class segmentation. Tawsif Sarwar, (2020) **"Phishing Attack Detection using Machine Learning Classification Techniques"** [4] The Phishing Detection method intended to predict whether web harassment or not using supervised ML methods and analyzing the URLs, website structure, and other distinguishing factors between phishing and lawful websites. Phishing Attack Detection combines a variety of supervised ML methods to categorize phishing and legitimate websites. Finally, their performance is evaluated and factored in to determine which of the supervised machine learning methods discussed is the most successful for the task at hand. The performance of three frequently used machine learning classifiers is compared in this research. The random forest has the best performance of the three classifiers, with a precision of 97 percent. The random forest's AUC is 1.0, indicating that our system can detect phishing websites with great accuracy. Da Xiao; Meiyi Jiang, (2020) **"A Malicious Mail Filtering and Tracing System based on KNN and an Improved LSTM Algorithm"** [5] Malicious Mail Filtering describes a phishing email filtering system that uses machine learning and deep learning. Not only can the system distinguish between good and bad emails (such as spam and phishing emails), but it can also identify the source of phishing emails based on the similarity of emails sent by the same attacker. The KNN and Bi-LSTM-Attention classifiers are used extensively in the system. The categorization result is relatively good, indicating that it is worthy of future study and implementation. Rabiei Mamat, (2019) "**E-mail Spam Detection Simulated Annealing using a Hybrid Water Cycle Optimization Algorithm"** [6] The dimensionality of spam email classifiers is the topic of this research. As a result, the feature selection method might be a curse for the dimensionality of relevant feature selection and classification. However, by removing and reducing redundancy several of the features can be reduced, and training time can be raised. The classification performance will improve as a result of this. The drawbacks of the popular algorithms employed in prior feature selection studies were explored in this paper. Using seven datasets, three common classical classifiers were highlighted: KNN, NB, and SVM. Examining the proposed WCFS, compared it to three prominent feature selection methods, and evaluated it in terms of performance (GA, PSO, and HS). Yong Fang; Cheng Zhang, (2019) **"Improved RCNN Model for Phishing Email Detection with Multilevel Vectors and Attention Mechanism"** [7] The author of this report began by examining the structure of emails. The researchers later introduced THEMIS, a novel fraudulent mail prediction system based on enhanced Recurrent CNN models (RCNN) with multidimensional dimensions and an attention mechanism. THEMIS is used to mimic mails at the header, body, character, and word levels all at once. The THEMIS model yields an encouraging result. The benefits of the proposed THEMIS model are demonstrated through a series of experiments. Tommy Chin, (2018) **"Phishlimiter: A Phishing Detection and Mitigation Software-Defined Networking Approach"** [8] PhishLimter is an innovative analysis and prevention strategy proposed by the study's author, in which system provide a new SPI technology and integrate it with SDN to detect phishing activities via e-mail. There are two parts to the suggested DPI solution: phishing signature categorization and real-time SPI. Based on each deep packet inspection, assessing the dependability of each SDN flow to identify any potential dangers Similarly, noticed how PhishLimiter's suggested inspection strategy, which includes two SF and FI modes, detects and mitigates phishing assaults before they reach the end-users if the flow is deemed untrustworthy. Reshma Varghese,

(2017) "**Feature Set for Effective Spam Email Filtering"** [9] The Spam Email Filtering contributes in three ways: Finding an efficient training dataset for junk mail filtering, analyzing the four types of features, including BoWs, PoS Tag, analyzing using the four types of features, including BoWs, Bigram Bag-of-Word (B-BoW) Investigated a model's performance on individual features as well as the aggregation of optimal features; Investigated a model's performance on classifiers such as AdaBoostJ48, Random Forest, and SMO; Investigated a model's performance on particular components and also the aggregation of optimal features. To determine the optimal feature set for spam email filtering, different kinds of characteristics were retrieved from the EnronSpam dataset. This study primarily used features from four categories: Bag-of-Word (BoW)s, Bigram Bag-of-Words, PoS Tag, and Bigram PoS Tag features. Simranjit Kaur Tuteja, (2016) "**Spam filtering via email using the BPNN classification algorithm"** [10] A neural network's basic structure is made up of input and output layers, with hidden layers in between that have fluctuating or constant numbers of neurons. The neurons operate as simple computing units in the training phase, simulating biological signal propagation and minimizing the experimental risk. The use of ANN Feed Forward and Back Propagation, as well as the k-mean clustering technique in the pre-processing stage, will improve the efficiency of spam and phish identification and filtration. posterior odds ratio (Zellner, 1979) for comparison of these Models.

III. EXISTING SYSTEM

The Filtering Techniques Based on Content: Algorithms examine terms, their frequency, and the propagation of words and expressions within the text to classify e-mails as spam or non-spam. **The technique of Spam Detection Based on Cases:** Methods educated on well-annotated malicious mail and anti-malicious mail divide incoming emails into two categories. **Spam Filtering Techniques Using Heuristics or Rules:** The algorithm utilizes pre-defined rules imposed as regular expressions to assign a score to e-mail communications. They categorize emails into spam and non-spam subcategories based on the scores they receive. **The Previous Spam Filtering Method Based on Likeness:** By constructing a drawing point for each new occurrence and vector in a latent space, algorithms extract the properties of incoming messages. These new points are assigned to the nearest non-spam and spam classifications using the KNN algorithm. **Technique for Adaptive Spam Filtering:** Algorithms that classify emails into distinct categories based on comparison scores between each group and a set of pre-defined groups distinguish non-spam and spam emails. **Classifiers based on machine learning:** Models for machine learning are chosen for their diversity, acceptability, and inclusion in the machine learning community. SVM, Naive Bayes (NB), and Deep Learning are various types of classifiers (DT). The disadvantages are: Current email classification methods result in the loss of critical data and/or irrelevant emails. The block listing procedure is primarily concerned with identifying and documenting people, which takes a substantial amount of time and effort. For the representation of features that aren't well-suited to machine learning approaches, manual feature engineering is required. Because forensic software relies on keyword searches, it frequently generates irrelevant e-mails.

IV. PROPOSED SYSTEM

Data collection, pre-processing, feature extraction, parameter tuning, and classification using the LSTM-GRU model are all part of the proposed technique. E-mail datasets are classified into four categories in this analysis: normal, disturbing, suspicious, and fraudulent. The email's body is broken down into word levels, and the embedding layer is used to train and extract the vector sequence.

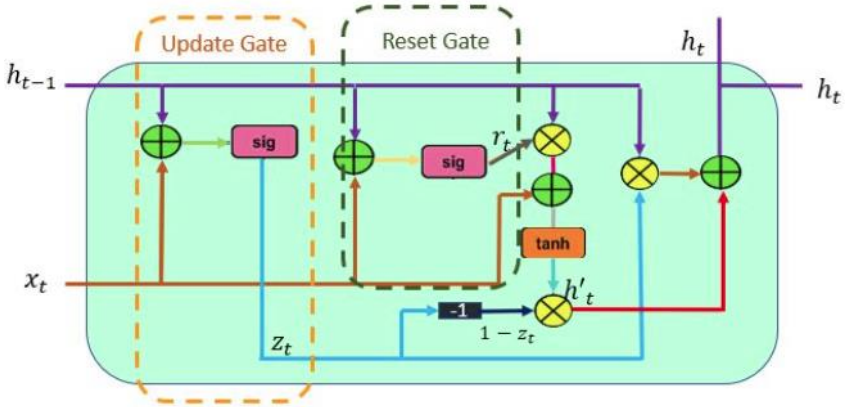

Figure 3: Update Gate and Reset Gate

The advantages are: verify the accuracy of e-mails effectively. Maintaining the categorization process's sturdiness and reliability. Spoof diagnosis is aided by e-mail content analysis. Branding does not necessitate the use of human resources. Scrutinize the e-mail server for potentially hazardous or unwanted messages.

V.SYSTEM DESIGN

A System design is a design that is used to abstract the overall outline of the software system and the relationships, constraints, and boundaries between components. It is an important tool as it provides an overall view of the physical deployment of the software system and its evolution roadmap.

System Architecture – Training Phase

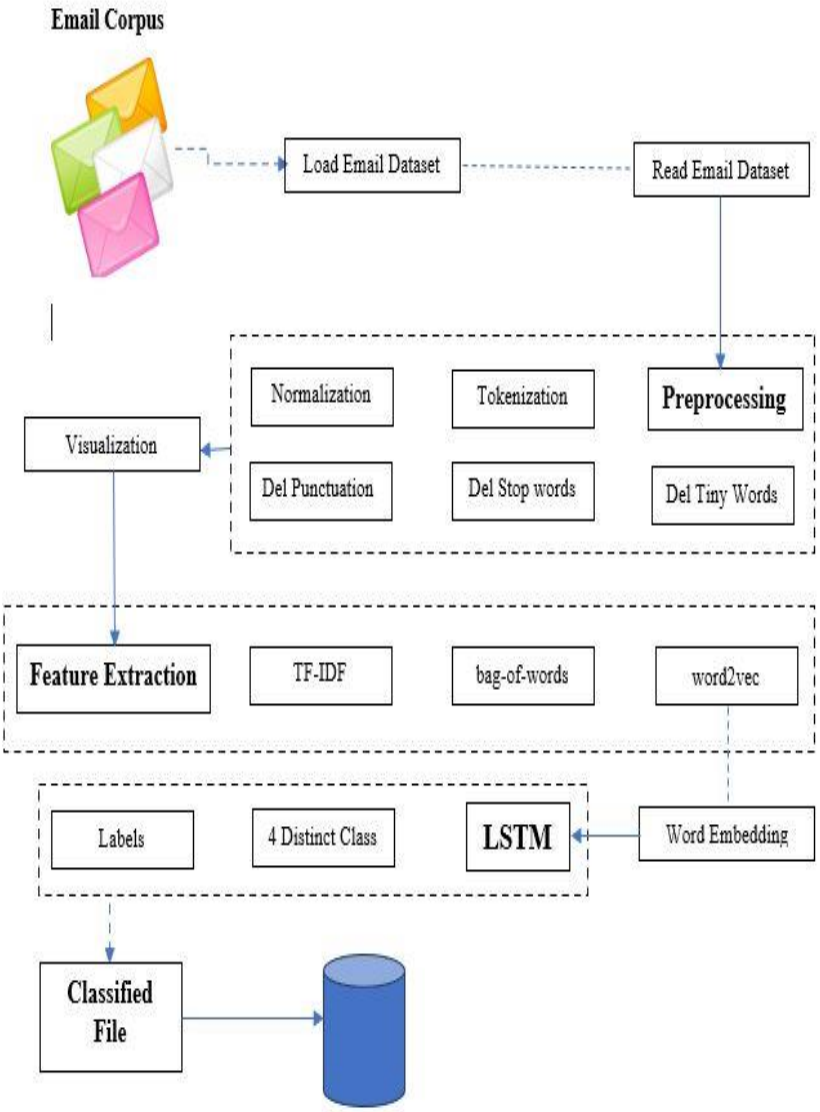

Figure 4: Training Phase

System Architecture – Testing Phase

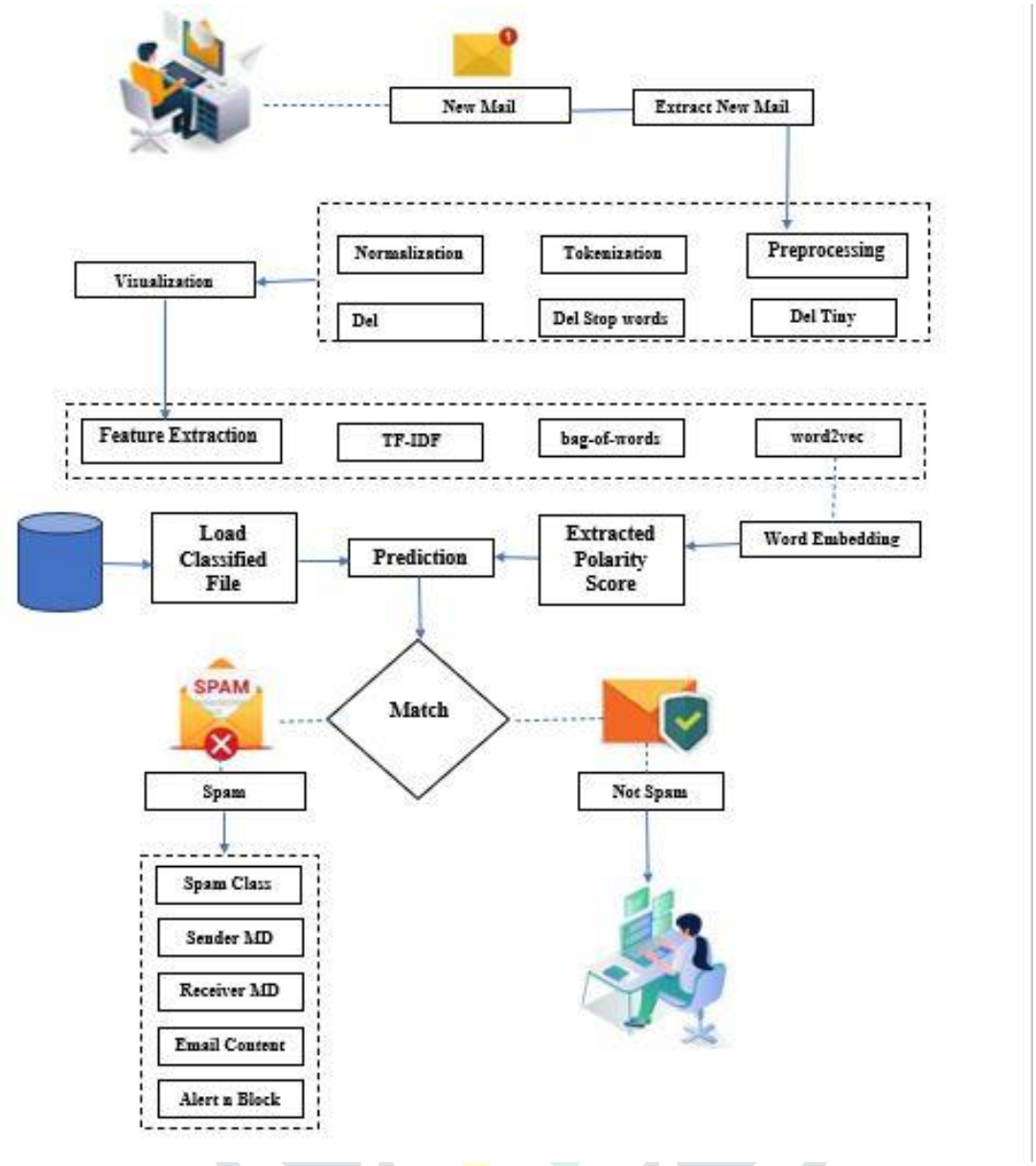

Figure 5: Testing phase

VI. IMPLEMENTATION

6.1 DATA PRE-PROCESSING

The statistics pre-processing segment incorporates sports that use herbal language to standardize and put together the textual content for analysis.

6.1.1) TOKENIZATION: Tokenization is a degree in herbal language processing wherein the uncooked textual content is damaged down into its constituent pieces. The approach of reworking texts into phrases with the usage of hard and fast regulations is called tokenization. The tokenization approach has performed the usage of Python's SpaCy package.

6.1.2) STOP WORDS REMOVAL: Words like "a" and "the," that are generally utilized in e-mails, are superfluous and upload noise to the textual content data. Stop phrases are sentences that may be eliminated from textual content previous to it being processed. Stop phrases have been eliminated from the textual content the usage of the Python program "NLTK".

6.2.3) PUNCTUATION REMOVAL: Punctuation (e.g., complete stop (.), comma (,), brackets) is used to interrupt statements and make clear meaning. To remove punctuation, the "NLTK" library is used.

6.2 FEATURE EXTRACTION

After the needless gadgets are eliminated, the prolonged listing of phrases is transformed into numbers. The TF-IDF technique becomes used to complete this challenge. The period frequency is the variety of instances a phrase seems in a document, and the IDF is the ratio of the whole variety of files to the variety of files containing the period. The textual content bag-of-phrases version is a truthful and extensively used method for extracting statistics from
textual content.

$$TFIDF = (fl * (df)$$

$$TFIDF = tf * lnverse(df)$$

$$TFIDF(t,d,D) = TF(t,d).IDF(t,D)$$

$$TFIDF(t.d) = \log \frac{N}{1\ di:Dti:DI} \quad \text{.....(1)}$$

It's additionally important to rent DL to extract traits from a word's context. This became done through the use of a word2vec neural network-primarily based technique. The following equation demonstrates how word2vec governs the word-context use of opportunity measures. (w; c) is a word-context pair selected from the big set D, and D is a pair-clever illustration of a group of words.

$$p_{(D=1|w,c,k)} = \frac{1}{1+e^{(W\cdot C1+w\cdot ci+\ldots+w-q)}} \quad \ldots\ldots\ldots\ldots(2)$$

The multi-phrase context is likewise a type of word2vec, as tested in Equation 2. The variable-period context is also managed via way of means of the arithmetic proven below.

$$P_{(D=1|w.c)} = \frac{1}{1+e^{-s(w,c)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

6.3 Word Embedding Layer

The approach of translating phrases into real numbers is referred to as embedding. Many system mastering and deep mastering algorithms can best study from numerical values and can not recognize entries that have now no longer been processed (textual content form). The approach of phrase embedding is used to organize textual content-to-numbers conversions. It extracts textual content attributes as real-valued numbers and organizes them as such. A phrase mapping dictionary is used to transform the phrases (phrases) right into an actual fee vector. Machine studying function engineering procedures have essential flaws: the primary is that statistics are represented by the usage of sparse vectors, and the second one is that phrase which means is neglected. In the embedding vectors, comparable phrases can be represented through honestly actual-valued values. The phrases love and affection, for example, can be near each other inside by considering vector.

6.4 LSTM BASED GRU CLASSIFICATION MODEL

E-mails are labeled as Normal, Harassing, Suspicious, or Fraudulent via way of means of LSTM. We included the LSTM and the GRU to make use of their gated architectures due to the fact they're each primarily based totally on a gated community design.

The layered shape of DL fashions allows self-mastering in terms of ML version deployment. A variety of libraries offer a complete mastering implementation framework. We separated the information into 3 training, validation, and checking out units the usage of a 65:10:25 ratio. We extracted traits from email textual information on the usage of the phrase embedding approach. Using the one-heat encoding method, we divide the aim information into 4 groups. For excellent email types, we ship all pre-processed information to the novel shape of LSTM layer versions. We appoint LSTM layers with special GRU and Convo1D layer variations to show the entered textual statistics into a powerful email type system.

Another method for figuring out word vector similarity is Jaccard similarity, that's described by the use of Equation 4.

$$simJacard(u,v) = \frac{L;min(u[i]\bullet V[iJ)}{L;max(ulil\cdot V[iJ)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(4)$$

6.5 EVALUATION AND RESULTS

A classifier's common standard overall performance has measured the use of quite a few measures. The confusion matrix, which includes 4 terms, is used to organize those measurements.

- True positive (TP) values are people who can be said in this manner.
- True Negative (TN): These are values that have been specific as bad.
- False Positive (FP): bad effects are mistakenly labeled as positive.
- False Negative (FN): information that became incorrectly categorized as bad.

The following metrics are used to evaluate the overall performance of our proposed model.

A. ACCURACY

Is the utility fraction as a percentage of general programs accurately calculated? The accuracy of a detection technique can be calculated with the use of an equation.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(5)$$

B. PRECISION

Are the predicted programs in terms of the entire range of applicants certainly encouraging? Equation 6 can be used to compute it.

$$Precision = \frac{TP}{TP+FP} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

C. RECALL

By multiplying the anticipated well-categorized programs with the aid of using the whole quantity of successfully or incorrectly categorized programs, the remember is calculated. Recall can be calculated with the use of Equation 7.

$$Recall = \frac{TP}{TP+FN} \quad \ldots\ldots\ldots\ldots\ldots\ldots(7)$$

D. F-SCORE

The harmonic imply of accuracy and taken into account is the F-rating. It refers back to the model's capability to differentiate among little features. Equation 8 can be used to get the f-rating of a detection model.

$$F-score = 2 * (Precision.*Recall) / (Precision + Recall) \quad \ldots\ldots\ldots(8)$$

VII.EXPECTED OUTCOME

In Figure 5, the text is analysed using LSTM for Natural Language Processing, and the minimum-maximum mean result is calculated for both the Normal and Fraudulent labeled classes. The four mentioned classes represent the best-trained model.

FIGURE 6 depicts how the novel combination of LSTM and GRU improved the forensic analysis accuracy of a large sequence of E-mails datasets. The maximum length of an e-mail is over 1000 words, necessitating the employment of many sequence learning modules; typical sequence learning techniques include the LSTM and GRU. To show sequence learning, we used the E-mail dataset. The LSTM model has the second-best accuracy, while GRU has the worst accuracy for the E-mail dataset. Individual Deep Learning models, such as the LSTM and GRU, make predictions about the model's accuracy.

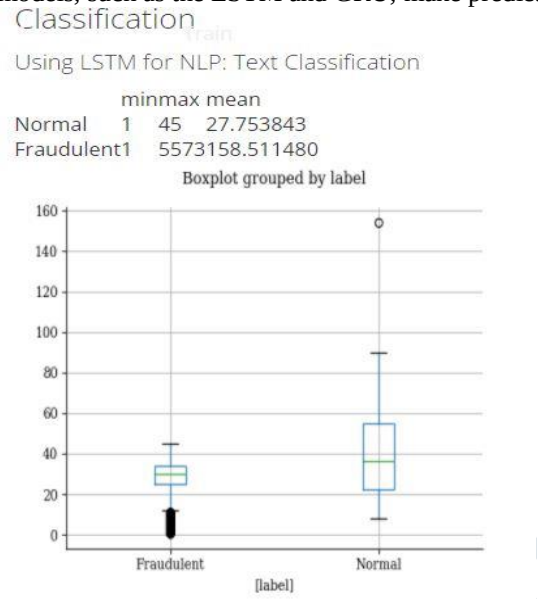

Figure 5: Normal Vs Fraudulent

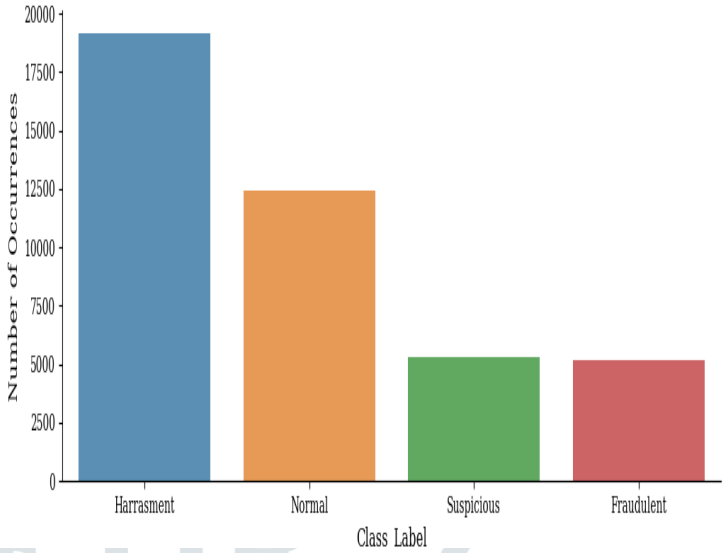

Figure 6: Four labeled classes

VIII. CONCLUSION

The proposed system is an LSTM model with an E-MailSinkAPI: Deep Learning Multiclass E-Mail Classification Model for Forensic Analysis We evaluated the proposed E-MailSinkAPI model using evaluation metrics such as precision, recall, accuracy, and f-score. Experimental results revealed that E-MailSinkAPI performed better than existing ML algorithms and achieved a classification accuracy of 95% using the novel technique of LSTM with recurrent gradient units. As different types of topics are discussed in E-mail content analysis. Many criminal activities are also performed through E-mails, but the E-mail repository is not available for public usage for privacy and security reasons. The non-availability of datasets on negative topics is a big hurdle in this research domain. Many researchers had just mentioned reports about criminal activities performed by E-mails, but they could not experiment due to a lack of datasets. For now, we are considering e-mail classes such as normal, harassment, fraudulent, and suspicious; however, many other r classes can be added to this work in the presence of the massive amount of e-mail data. We intend to produce datasets on these topics and build a generalized model for E-mail classification in the future.

REFERENCES

[1] Sanaa A. A. Ghaleb, "Enhance Grasshopper Optimization Algorithm for Spam Detection System Training Neural Networks", 2021.
[2] Ivana Čavor, "Classifying Emails Using a Decision Tree Model", 2021.
[3] Luis Felipe Gutiérrez, "Phishing Detection Using Email Embeddings", 2021.
[4] Tawsif Sarwar, "Phishing Attack Detection using Machine Learning Classification Techniques", 2020.
[5] Da Xiao; Meiyi Jiang, "A Malicious Mail Filtering and Tracing System based on KNN and an Improved LSTM Algorithm", 2020.
[6] Rabiei Mamat, "E-mail Spam Detection Simulated Annealing using a Hybrid Water Cycle Optimization Algorithm", 2019.
[7] Yong Fang; Cheng Zhang, "Improved RCNN Model for Phishing Email Detection with Multilevel Vectors and Attention Mechanism", 2019.
[8] Tommy Chin, "Phishlimiter: A Phishing Detection and Mitigation Software-Defined Networking Approach", 2018.
[9] Reshma Varghese, "Feature Set for Effective Spam Email Filtering", 2017.
[10] Simranjit Kaur Tuteja, "Spam filtering via email using the BPNN classification algorithm", 2016.