Transfer Learning-based Forensic Analysis and Classification of E-Mail Content

FARKHUND IQBAL, Zayed University, UAE
ABDUL REHMAN JAVED, Lebanese American University, Lebanon
RUTVIJ H. JHAVERI*, Pandit Deendayal Energy University, India
AHMAD ALMADHOR, Jouf University, Saudi Arabia
UMAR FAROOQ, National University of Computer and Emerging Sciences, Pakistan

Emails have become a crucial element in societal transformation recently. Spam emails have also become more common, making spam filters more important. Several approaches have been attempted to classify emails as spam or non-spam based on their content. However, it is necessary to classify emails based on their contents rather than just analyzing titles, links, and URLs. This paper proposes a novel approach for email analysis and classification based on content into four categories: Normal, Fraudulent, Threatening, and Suspicious emails. We explore the challenges and issues in this forensics paradigm and present a transfer learning-based approach to email forensics. We demonstrate several use cases where email forensics can provide crucial insights for discovering information about criminal situations. Machine learning (ML) and deep learning (DL)-based approaches are used for helping email forensics investigations. We conducted a detailed analysis of ML/DL and transformers to identify the best email analysis approach. It is observed that transformers-based architectures, such as Bidirectional Encoder Representations from Transformers (BERT), achieve the best accuracy of 98%, surpassing existing studies. Furthermore, we discuss various challenges and limitations associated with email forensics, such as data privacy concerns and the need to continuously update the classification model to keep up with new spam techniques.

CCS Concepts: • **Computing methodologies** → **Language resources**.

Additional Key Words and Phrases: Multiclass E-mail Classification, Digital Forensics, Artificial Intelligence, Fraud Email Detection, Threatening Email Detection, Suspicious Email

1 INTRODUCTION

Email has been a dependable, secure, and official communication medium. People's dependence on email has increased as people have gained confidence in its reliability and security. As the economy has expanded, the nature of attacks has shifted dramatically, moving away from being random and generic and becoming very strategic and intricate. Unsolicited email, sometimes known as spam, is a source of various cybercrime strategies that employ

*CORRESPONDENCE: rutvij.jhaveri@sot.pdpu.ac.in

Authors' addresses: FARKHUND IQBAL, farkhund.Iqbal@zu.ac.ae, College of Technological Innovation, Zayed University, Abu Dhabi, UAE; Abdul Rehman Javed, abdulrehman.cs@au.edu.pk, Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon; Rutvij H. Jhaveri*, rutvij.jhaveri@sot.pdpu.ac.in, Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, India; AHMAD ALMADHOR, aaalmadhor@ju.edu.sa, Department of Computer Engineering and Networks, College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia; Umar Farooq, umar5555f@gmail.com, Department of Computer Science, National University of Computer and Emerging Sciences, Islamabad, Pakistan.

2375-4699/2023/6-ART $15.00
https://doi.org/10.1145/3604592

complex methods to deceive targeted victims [30, 31, 52]. Spam detection has emerged as one of the most vital applications of text forensics in the recent decade. However, fine-grained spam detection into subclasses should be prioritized. The massive increase in email data has made email management quite complex [71]. Emails may be sorted and classified depending on the sender, website name, length, and time [42, 49, 72]. However, identifying and analyzing emails based on their contents is required.

In the past, several methods for classifying emails as spam or non-spam based on their content were used. This study uses a multiclass email categorization technique to set up an orderly inbox. Forensics plays a critical role in modern law enforcement and criminal investigations, email system security and is also used in civil cases, corporate investigations, and other fields [18]. Forensics aims to provide objective, scientifically-based information and analysis that can be used in legal proceedings.

The analysis of electronic communication, such as emails, is essential to digital forensics, as it can provide critical evidence in legal investigations. However, the vast amount of email data makes it challenging to sort and analyze messages efficiently. As a result, there is a growing need for automated email categorization techniques that can accurately identify and classify different types of emails based on their content. This study addresses this need by proposing a multiclass email categorization technique by combining deep semantic analysis with machine learning and deep learning; we identified salient features of email content and assigned them to one of four categories: Typical, Fraudulent, Dangerous, or Suspicious. Semantic analysis has been mostly ignored in email forensics and categorization until recently. For several NLP tasks, including topic labeling, sentiment analysis, and language translation, deep learning has become increasingly popular [6, 13, 51]. Text representation [10, 33] is an important issue in natural language processing. Further, deep learning has recently been widely applied in various natural languages processing tasks such as subject categorization, sentiment analysis, and language translation. Research directions have changed due to the success of transformer models and transfer learning in natural language processing [6, 17, 27, 28]. In this research, we employed BERT embedding with and without word phrasing for deep semantic analysis and machine learning and deep learning techniques to categorize emails. Existing email categorization methods result in irrelevant emails and/or the loss of valuable information.

Motivation: Initially, spam was just used to attract commercial interests, such as URLs leading to company websites, but as time went on, scams, threats, and harassment via email became quite widespread. There are presently various approaches for identifying spam. However, they rely on more information, such as DNS addresses, the frequency with which a source sends emails, bulk email analysis, and email headlines. With the methodologies above, several researchers applied machine learning to detect spam [29, 31, 50] but they did not focus on approaches based on transformer topologies that manage not just short sequences but also extensive dependencies of large characters. As a result, we are interested in email forensic approaches based on pure text analysis that detect unsolicited emails based on their content.

This research makes the following contributions:

- Propose a transfer learning-based approach on SeFACED[1] dataset, an innovative and fast approach for categorizing E-mails into four separate classes: Normal, Fraudulent, Threatening, and Suspicious E-mails, based on Transformer topologies that manage not just short sequences but also extensive dependencies of large characters.
- BERT and its variants effectively extract useful information from emails that can be utilized as evidence in forensic investigations. E-mail content analysis aids spoof detection since analyzing the headers of individual emails rather than all emails is more efficient.
- Compare the performance of our transformer-based technique to the findings of SeFACED, classic ML and DL models, and previous research on E-mail content analysis and categorization. We achieved 96% accuracy

[1] https://github.com/Abdul-Rehman-J/SeFACED

and 95.8% F1-score because of deep feature engineering. Our deep learning and transformer-based models are 98% accurate.
- The findings show that the pretrained models efficiently classify E-mail content with a 98.0% accuracy, a precision of 97.8%, a recall of 98.2%, and an F1-score of 97.99%, maintaining the classification process robust and dependable.

The rest of the paper is organized as follows. Section 2 provides the related work on forensics analysis and email content classifications. Section 3 provide the proposed approach. Section 4 provides the results and discussion. Section 5 provides the ablation study. Section 6 concludes this paper.

2 RELATED WORK

Information security businesses have created various computer forensic programs [1, 2, 34–36]. These tools concentrate on important functions, including email data collection, decoding, and developing fundamental network graphs [4, 7, 36]. Unsupervised or rule-based work on email classification, spoofing, and phishing is included in supervised ML and DL work. In each area, the subsections detail the various ways of processing E-mail data:

2.1 Rule Based Detection

Rule-based strategies collect information from text by classifying it into various groups using handmade linguistic rules and fuzzy logic or some established rules. Based on its content, the input text is categorized using semantic parameters. Certain words assist the investigator in determining if a text is normal. Spam or fraudulent writing has a few distinguishing words that assist in distinguishing it from normal English. When the quantity of regular words in the email surpasses the number of spam or fraudulent phrases, the email is considered normal. They work by enforcing a series of predefined rules, each with its weight. The spam, fraudulent, and harassing corpora are examined for harming material, and any techniques identified in the text are given some extra weighted value in the total score [26, 45, 59, 66]. Nadja et al. [53] presented a hierarchical semantic analysis of emails to detect domain-specific irrelevant messages. The starting level seeks to categorize emails to create a domain-specific irrelevant content categorization. The last level tried to extract semantic information in each domain as rules that differentiate between spam and genuine communications.

Based on the prior studies on email categorization using rule-based approaches, we can infer that researchers value rule-based techniques for their usefulness in email forensics. Email forensic researchers favor freely available libraries and packages that assist in formulating specific techniques for various types of irrelevant content. Some rule-based techniques rely on fixed techniques that cannot be altered, making them incapable of dealing with continually changing content. The set techniques must be improved daily to increase the method's capacity for better text analysis. Automated rule generation can be employed to deal with the variable nature. Rule-based systems have substantial downsides to complex systems regarding computational cost, analytical vagueness, and rule architecture.

When it comes to the complex information management requirements of Intelligent Transportation Systems (ITS), Wireless Sensor Networks (WSN) play a crucial role. Due to the strict requirements for security and dependability while collecting data from installed sensor nodes, the suggested framework (ICITS) mainly aims toward road transport in military regions. By comparing ICITS' simulation results with those of the newly proposed Cluster-based Intelligent Routing Protocol, this work finds that it outperforms the latter by%ages of 54.7%, 19.6%, and 40.5%, respectively, on a variety of performance criteria [68].

2.2 Machine Learning-Based Detection

Different machine learning algorithms (i.e., unsupervised and probabilistic) have been used to detect spam content. Machine learning is divided into two types: supervised machine learning, which requires previous knowledge of

the target variable, and unsupervised machine learning, which requires no prior knowledge of the target variable. Both are extensively used methodologies in Natural Language applications. To overcome the limitations of existing filtering methods, Jancy et al. [22] employed the Bayes approach and the Markov Random Field. The system built using these two models was efficient in execution, i.e., reduced execution time, increased accuracy, and successfully combined the best practices of both models.

The study gives an overview of digital forensics, the issues that investigators confront, and how machine learning techniques can help solve these problems [8]. The authors discuss the potential advantages of machine learning in digital forensics, such as improved accuracy and efficiency in finding and processing digital evidence. The authors also address decision trees, support vector machines, and neural networks as machine learning approaches that can be employed in digital forensics. Lastly, the report emphasizes the importance of more research in this field to advance the application of machine learning in digital forensics.

The research by Ahmed et al., [5] discusses the use of machine learning techniques for spam detection in email and Internet of Things (IoT) platforms. The authors comprehensively analyze various machine learning algorithms used for spam detection, such as Naïve Bayes, Decision Trees, Support Vector Machines (SVM), and Random Forest. The study also highlights the challenges and limitations of using machine learning for spam detection, including issues with feature selection, data imbalance, and the need to update the model continuously. The authors conducted experiments on two datasets, namely Enron Email Dataset and IoT Email Dataset, to evaluate the performance of different machine learning algorithms. The results show that SVM and Random Forest algorithms outperform the other algorithms in accuracy, precision, recall, and F1-score. The authors also discuss future research directions, such as integrating deep learning and natural language processing techniques to improve spam detection in emails and IoT platforms. The study highlights the challenges and limitations of using machine learning for spam detection and presents numerical results from experiments conducted on two datasets.

Dedeturk et al. [23] compared their proposed email analysis strategy against that of commonly used models. They ran their system through text filtering tests and e-mail datasets that are publicly available and observed that its performance was much better than its counterparts. To identify the appropriate category of e-mails, Nayak et al. [50] used a hybrid technique that included Naive Bayes and Decision Tree algorithms (DT). Using their hybrid technique, an accuracy of 88.12% was achieved by them. Sharma et al. [58] employed DT and K-NN classifiers to protect consumers from fraud. They put their strategy to the test UCI e-mail dataset. For their classification model, the Decision Tree classifier generated superior results, with accuracy and F1-score of 90 & 91.5%, respectively. Authors in [47] discovered that multi-algorithms easily beat single-algorithm in email forensics. They examined the performance of both supervised and unsupervised models in terms of accuracy and detecting e-mails as normal or spam. The supervised strategy outperformed the unsupervised approach in terms of spam detection.

Junnarkar et al. [38] used a two-pronged approach to prevent spam from reaching his email contacts. The proposed solution is to monitor the URL to determine whether or not the links in the email were broken or fake. Five different ML algorithms were examined. NB and SVM got more than 90% accuracy on the e-mail dataset. Researchers investigate the relevance of ML approaches for spam text categorization in their study, concluding that ML techniques overcome the shortcomings of rule-based strategies for this detection [9, 60, 63]. An ensemble method may combine multiple ML classifiers with improving text classification tasks. ML techniques' computational complexity and domain dependency have been highlighted in studies of ML algorithms for Email forensics. Due to the lengthy training time and large datasets required by many ML techniques, the researchers suggest DL methods to alleviate these limitations in email forensics.

2.3 Hybrid Approaches

Combining an ML/DL-based classifier with specific stated fuzzy logic rules, hybrid e-mail forensics systems can better classify spam. To detect spam in emails, Kaddoura et al. [39] employed a unique technique that is comprised

of "Rule-Based Subject Analysis" (RBSA) with different ML/DL algorithms. Their proposed method involves assigning appropriate weights to fraudulent content and constructing a matrix, which is input into a classifier. Venkatraman et al. [67] used the Spambase and Enron corpus datasets to evaluate their hybrid spam categorization approach. They have a near-perfect accuracy rate of 98%. In their study, Wu et al. [70] used a novel technique that combined Neural Networks (NN) with RB algorithms. To categorize spam material, they employed NN, RB pre-processing, and behavior detection modules using a new approach that encodes all these techniques together. They tested their strategy on a million-email corpus and achieved a 99.60% accuracy rate.

2.4 Deep Learning based approaches

Because of their capacity to handle difficult problems, DL models are gaining appeal among NLP experts. DL is based on training an extensive neural network inspired by brain processes utilizing enormous amounts of data [41, 65]. They can deal with scalability concerns and extract data attributes automatically. Deep networks like CNN and LSTM networks are the most popular DL models among NLP researchers. CNN, one of the most significant and widely used DL algorithms, has recently attracted much interest as a solution to NLP problems. It has been effectively utilized in sentiment analysis, [40], picture [57] and text classification [61], pattern recognition [48], and a variety of other applications.

Lai et al. [43] employed recurrent units for data, information, and context mining from the text for classification purposes. They outperformed convolutions networks in text classification and semantic information extraction tasks. Tai et al. [62] use Recurrent units and LSTM to extract the sequential information from the text, and they stated that LSTM and recurrent units perform best for most textual tasks. Apart from text, RNN/LSTM also performs best for the time series data, where data is sequential. These types of networks are widely used in many NLP applications. Bashar et al. [11] examined the 34,519 records in the Enron corpus with an LSTM network and a GRU model to identify spam. In a prior work inspired by LSTM and GRU's Hina et al. [29], fine-grained email forensics reached an accuracy of 96%. Tong et al. [64] used a DL approach based on LSTM and BERT to deal with issues including skewed representation, weak detection, and poor practicality in Chinese fraudulent email detection. A long-short attention mechanism was employed in developing this strategy for capturing complex text features. Liu et al. [44] used a hybrid model consisting of Convolution structure and Bi-LSTM to extract complete and relevant semantics from a document to identify hotel-related spam reviews. If they can get an F1 score of around 92.8, they may outperform conventional categorization methods. More studies [12, 20] have used DL algorithms for text analysis, which may collect contextual information from the text to identify spam.

3 EMAIL DETECTION AND CLASSIFICATION

This section details the experimental setup and implementation platforms we use to implement our proposed solution, as depicted in Figure 1. We design this pipeline in Python using modules like pandas, NumPy, TensorFlow, Keras, PyTorch, sklearn, and others. These libraries are widely used in Natural Language Processing (NLP) to accomplish tasks like classification, forensics, question answering, machine translation, and text summarization, to name a few. These libraries are the foundation for various computer vision, machine learning, and deep learning projects.

3.1 Dataset Selection

This study's dataset is a previously produced dataset that is now open to the public [29], a combination of four separate datasets. The dataset contains benign and malicious emails, such as those from the Enron Corpora [2], harassment messages from the Hate Speech and Offensive Speech datasets. The Email Forensics dataset enhanced data quality by including Twitter data and other dubious email sources. Terror-related tweets obtained using the

[2]https://www.cs.cmu.edu/ enron/

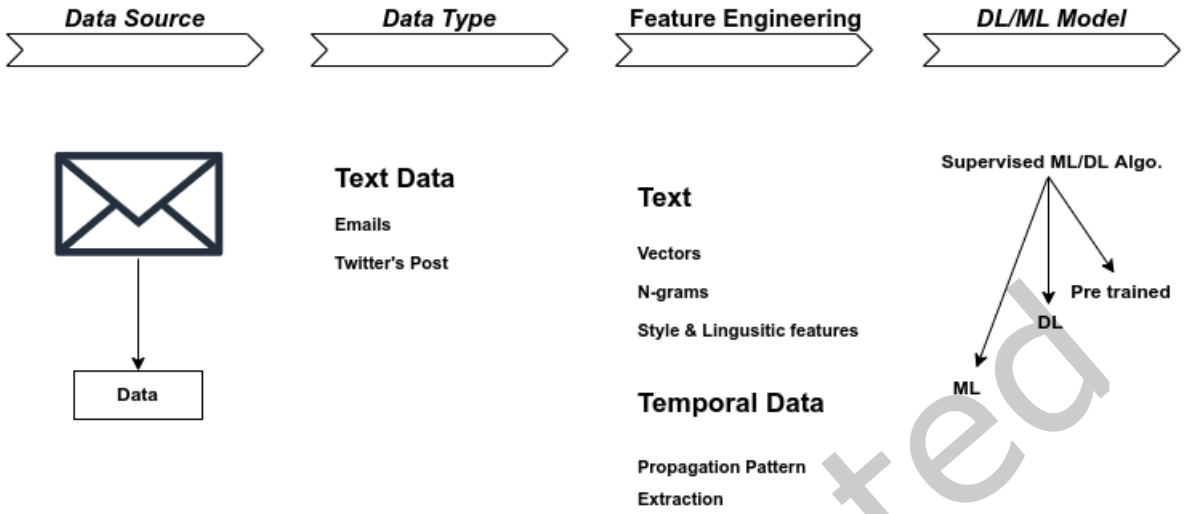


Fig. 1. Pipeline of complete processes from data collection and preprocessing to final model selection.

Twitter API are part of the suspicious dataset. Many datasets are combined into a single structured file to facilitate multiclass E-mail classification. Word Embedding, Word2Vector, and the Term Frequency-Inverse Document Frequency (TF-IDF) are used to classify the characteristics derived from the email content. The composition of the several E-mail corpora used in this study is detailed in table 1. Without the E-header mail's information—including the sender, subject, carbon copy recipients, and blind carbon copy recipients—the message's body text may be examined more thoroughly. Approximately 32,427 messages were included in the final dataset once it was constructed. The percentage represents the breakdown of the dataset into four categories. Consistent with other classes, ordinary email traffic (9001) accounts for 27.8% of the entire dataset. Table 1 demonstrates the balanced nature of the dataset.

Table 1. Percentage of different classes in the dataset

	Normal Emails	Fraudulent Emails	Harassment Emails	Suspicious Emails
Number	9001	9001	9138	5287
Percentage	27.8%	27.8%	28.2%	16.3%

3.1.1 Data Cleaning. Data cleaning/cleansing, integration, transformation, and reduction are the four processes in the data preparation pipeline. Figure 2 displays the steps involved in data cleaning pipeline:
Cleaning Data: The data is unstructured and includes URLs, emoticons, non-English words and phrases, emails, special characters, and blank spaces and lines. After eliminating any extraneous information from the data, a machine learning or deep learning model is trained. Figure 3 depicts the complete cleansing process.

Data Integration: Data integration is the second phase in the data preprocessing pipeline. It entails merging data from several sources. Databases, data cubes, and flat text files are examples of these sources. Meta-data analysis is also used to extract important information. Metadata can aid in the reduction of mistakes in schema

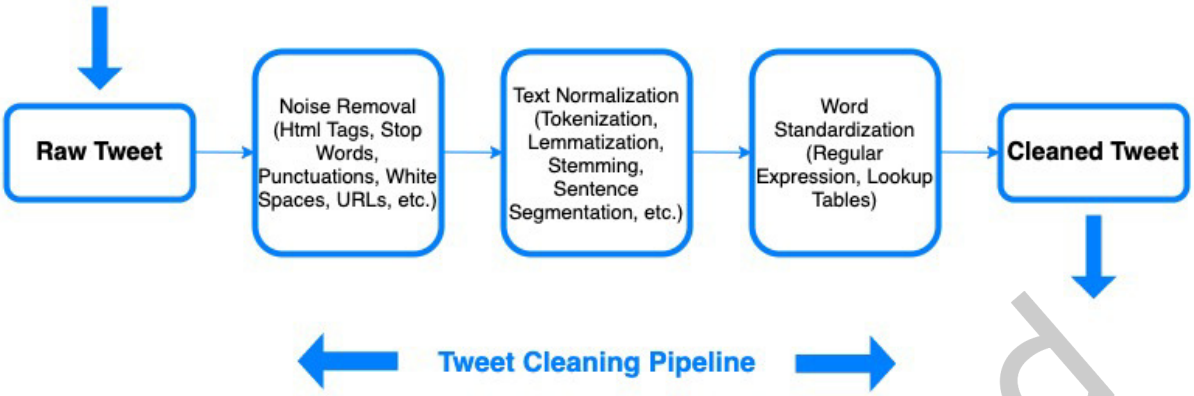


Fig. 2. Steps involved in data cleansing.

integration. Another critical issue is duplication. If an attribute is inherited from another table, it may be redundant. Inconsistencies in attribute or dimension names can also result in data set redundancy.

Data Transformation: Normalization, smoothing, aggregation, and generalization are all aspects of data transformation. These are the most often used strategies for normalizing data, i.e., between 0 and 1 or -1 and 1. Smoothing eliminates noisy data by defining upper and lower character boundaries.

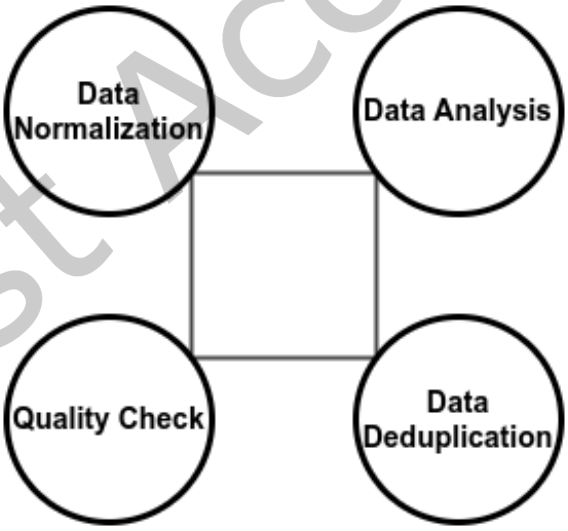


Fig. 3. Steps involved in data preprocessing and cleansing.

Data Reduction: Data reduction strategies include data compression and dimensionality reduction. This lowers the overall cost of training.

3.1.2 Dataset Samples. After preprocessing, cleaned data, i.e., removing all the extraneous information from the data to be supplied to the final model cleaning prior, would significantly improve performance: Table 2 contains the dataset samples:

Table 2. Dataset samples after preprocessing.

Tweet	Label
Greetings, Mr. Sir. Mr. Tambo is from Cape Town, South Africa, and I work for the South African Department of Mining and Natural Resources as an Executive Accountant. First and foremost, I used the word "remorse" to describe my regret . . .	Fraudulent
my parents want to go skiing from 12/09 - 12/14. Can you take off time then? the package is probably around $700-900/ Let me know -e	Normal
Era come so you can record him saying he's my bitch	Harassment
While Armenians have been given more than a week to leave our occupied territories, ethnic Azerbaijani Turks were . . .	Suspicious

3.2 Feature Engineering

Several feature engineering strategies for different features may be derived from the provided data. Some examples are word frequencies, i.e., TF-IDF, word embeddings, semantic features, linguistic characteristics, and so on. Word embeddings, such as Glove, FastText, Elmo, and BERT embeddings, are extensively used feature extraction approaches in text analysis applications.

3.3 Implementation Details

To discover the best-performing classifier for this job on our corpus, we would investigate a variety of machine learning classifiers and seq-2-seq models, i.e., BERT and its variants, as well as convolutional neural networks. It entails several phases: data collection, purification, preprocessing, rule-based approaches, and training ML/DL algorithms. Machine learning techniques are supervised and include regression and classification, which are used to train on data to predict, whereas deep learning approaches include simple artificial neural networks to advanced level seq-2-seq models like RNN and LSTM, as well as transformers like BERT and its variants. To reduce loss and obtain class probabilities, we employ cross-entropy loss. Figure 4 displays the complete system architecture from data acquisition to the final model.

3.3.1 Machine Learning Models. Naive Bayes (NB) algorithms that are Bernoulli (BNB) and Multinomial, Support Vector Machine (SVM) Classifiers, Regression Logistic and Linear, Random Forests (RF), Decision Tree (DT), and AdaBoost(AB) models will be used for training. Several ML classifiers should be explored to determine the best classifier for our corpus's unsolicited email detection job. Much of the research centers around some fundamental ML methods used to show the findings, such as Regression, SVMs, and Random Forests. We offered a three-pronged strategy with three distinct settings to identify the greatest potential answer to the stated challenge. With various machine learning models, we employ TF-IDF, CountVectorizer, and Word2Vec.

3.3.2 Vector Embeddings. We proposed a three-pronged strategy with three distinct settings to identify the greatest potential answer to the stated challenge. To begin, we utilize pre-trained FastText embeddings [15] and the HuggingFace tokenizer [69] to train these embeddings again on our corpus. The next step was to feed these previously trained embeddings of our corpus to perform classification into a FastText-like model using Keras' a Ktrain wrapper [46]. Second, we combine previously trained embedding vectors with the Bidirectional GRU model for classification. Finally, we apply and tweak two types of BERT to our corpus for classification. The first variation

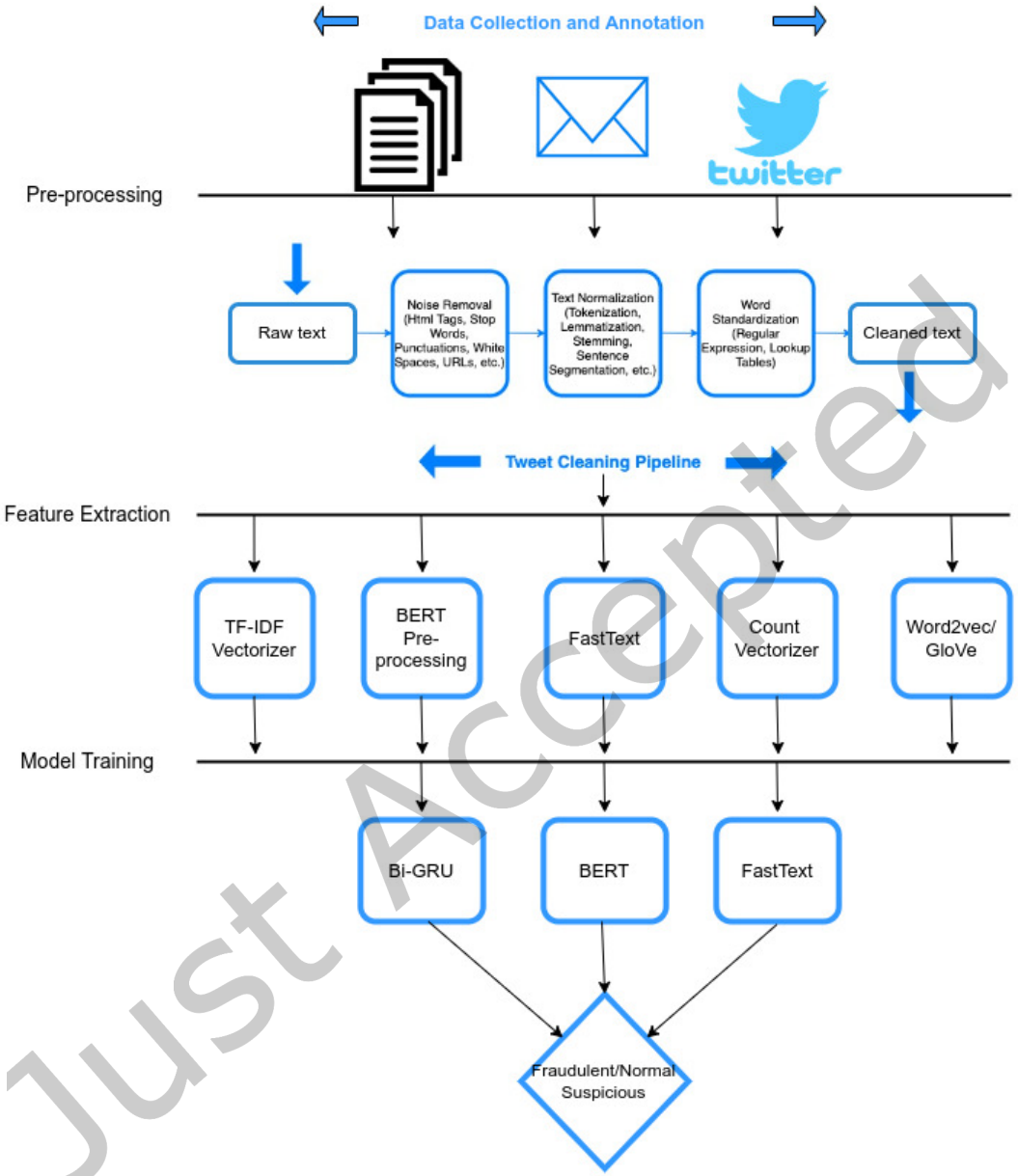


Fig. 4. Complete architecture of the proposed system, a multi-step pipeline that includes preprocessing, feature engineering, and model training.

involves freezing the whole BERT architecture and replacing it with a feed-forward network consisting of two dense layers & a softmax activation layer. The second method involves using transfer learning to train the whole BERT model on our dataset.

3.3.3 Deep Learning Models. However, as observed in the SemEval text-related contests, Many of the entries used Deep Learning approaches to classify text, such as CNN and LSTM [21], transformers, and Gated Recurrent Units (GRUs). The most important one is that, compared to state-of-the-art machine learning methods, neural network models perform better when utilizing pre-existing frameworks to extract information. Aside from these strategies, recent improvements have demonstrated that transformer-based attention models are currently considered cutting-edge models for many NLP applications [25]. Attention mechanisms may learn the context of words far more effectively and are more dependable in retaining contextual information.

3.4 Loss Function

It is critical to understand which type of loss function is appropriate for the presented issue to construct better machine/deep learning models for new ML/DL tasks [24]. For example, what distinguishes the chosen loss function from others? According to several research studies, cross-entropy is the most commonly utilized loss function in today's classification jobs for differentiating many classes [32, 56, 73]. Most deep learning models use the softmax activation function for classification to forecast probabilities and minimize cross-entropy loss [3]. Because our problem is likewise multiclass, we employ cross-entropy loss. Cross-entropy loss produces probability predictions, and the predicting class has a greater probability than the target class. A penalty is levied for incorrect predictions to minimize the loss value in the following epoch.

$$H(p,q) = -\sum_{n=1}^{N_c} p(x) log(q(x))$$

The above equation represents the cross-entropy loss function used in machine learning to evaluate the performance of a classification model. Here, the model attempts to predict the probability distribution of the true class labels p(x) given the input data and the predicted class labels q(x) outputted by the model. The sum is taken over all possible classes N_c in the problem. For a data collection of size N, a multiclass log loss cost function would be:

$$J = -\frac{1}{N}(\sum_{n=1}^{N_c} y_i log(\hat{y_i}))$$

In the above equation, we have a dataset of size N with N_c classes. The true label of the ith sample is represented by y_i. The log function is used to penalize the algorithm heavily for predictions far off the true value. If the predicted probability for the true class is high, the log value will be small and vice versa. The cost function J measures the average loss over the entire dataset. The negative sign in the equation indicates that we are minimizing the cost function.

3.5 Formal Solution Definition

- Find some often used terms in fraudulent, hate, and harassing emails by using lexicons and email datasets obtained via study of the topic

$$\boldsymbol{W} \in \{w_1, w_2, w_3, ..., w_n\}$$

- We have collected emails, messages and tweets

$$\boldsymbol{V} \in \{v_1, v_2, v_3, ..., v_m\}$$

  based on the words present in lexicon $\boldsymbol{W}$.
- Emails (v) that we collected classified that are given an email $\boldsymbol{V}$ a label

$$\boldsymbol{l} = \{\boldsymbol{n}, \boldsymbol{h}, \boldsymbol{f}, \boldsymbol{t}\}$$

- A corpus of email $\boldsymbol{V}$ messages has been compiled and their corresponding labels $\boldsymbol{l}$

$$\boldsymbol{V}, \boldsymbol{l} = \{(t_1, l), (t_2, l), (t_3, l), ..., (t_m, l)\}$$

- We then develop and implement a system to scrub tweets. Numerous widely-applied preprocessing methods, denoted by the letter D, are incorporated into this pipeline.
- Then we look for the word embeddings $\boldsymbol{E}$.

$$\boldsymbol{E}, \boldsymbol{V} = \{(e_1, v_1), (e_2, v_2), (e_3, v_3), ..., (e_m, v_m)\}$$

  As a consequence, we get a matched vector embedding for each email in corpus V. We may build a vector representation of $\boldsymbol{V}$ by merging these embeddings for each email v in $\boldsymbol{V}$.
- Finally, for training and prediction, these word embeddings $\boldsymbol{E}$ and actual labels are given to the simple ML/DL.

$$\boldsymbol{E}, \boldsymbol{l} = \{(e_1, l), (e_2, l), (e_3, l), ..., (e_m, l)\}$$

4 RESULTS AND DISCUSSION

In this part, we review the implementation details for the tests we ran for multiple baseline models and the best-performing model for detecting and analyzing emails. We also went through the in-depth embedding techniques employed. Then, using the previously indicated approaches, we analyze the experiment outcomes. We conducted various tests to confirm the superiority of our approaches to improve categorization and broaden prediction scope. According to the research, local contextual information adds little to the analysis. As a result, several words embedding approaches are helpful and can considerably improve the acquisition of contextual information *Benito et al.* [14]. We use FastText, BERT embeddings, and earlier word embedding techniques like Word2Vec, TF-IDF vectorizer, and CountVectorizer (CV) for existing approach models. Initially designed for the English language, they now support many other languages due to rapid research improvement and remarkable results. We use the tokenizer from the HuggingFace library for these models.

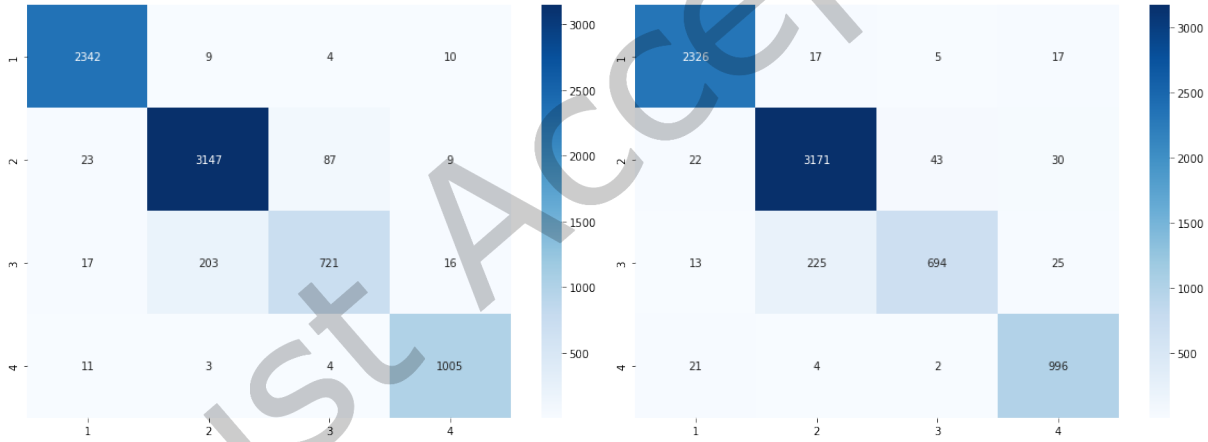
4.1 Baseline Results

We employed several baselines using our dataset to obtain baseline findings. We define tasks based on the corpus presented in the preceding section, which asks for class identification. The corpus was divided into three sections: training, development, and testing, with each section holding 70%, 20%, and 10% of the total. The entire system for misinformation detection consists of various machine learning classifiers with default settings, in addition to TF-IDF, CV, and Word2Vec. The length of the phrase is used to indicate the input occurrences. In addition to machine learning, we employ a variety of pre-trained and deep learning models as they can handle class imbalance efficiently, including BERT, xlm-Roberta, and distill-BERT, to leverage transfer learning for this classification job. Table 3 displays the performance of various baselines on the email forensics task. The baseline for the email classification challenge chooses one of the four classes.

KNN: is our initial baseline algorithm. It is the most apparent choice when considering a classification problem. (KNN baseline) with *(TF-IDF)*, Word2Vec 200d embeddings, and CV. We employ 1000d CV in conjunction with KNN with a *(RBF)* kernel and a *C* value of 100.
RF classifier: which also uses a 2000-dimensional TF-IDF vector representation, 200-dimensional Word2Vec embeddings trained on the dataset using 100 epochs and a 5-by-1 context window, and a CV and C value of 100 in sklearn, is our second baseline of choice [16].
SVM: Word2Vec 200d embeddings trained on the dataset for 100 epochs with a context window of 5 and a Support Vector Machine (SVM) with an RBF kernel and a C value of 100 make up the baseline models.
AdaBoost: is our fourth baseline option [55]. It uses the 2000d *(TF-IDF)* vector representation, 1000d CV feature

extraction method with AdaBoost with *rbf* kernel. The *C* value is 100. We use Word2Vec 200d embeddings.
Simple Perceptron: is the fifth choice for baseline. It employs the 2000d *(TF-IDF)* vector representation, sklearn's 1000d CountVectorizer, and Word2Vec 200d embeddings.
GaussianNB: is our sixth choice for baseline. It employs the 2000d *(TF-IDF)* vector representation, sklearn's 1000d CountVectorizer, and Word2Vec 200d embeddings. Our seventh baseline choice is QDA, which employs the 2000d *(TF-IDF)* vector representation. We additionally employ 1000d CountVectorizer and Word2Vec 200d embeddings that were trained on the dataset for 100 epochs with a context window of 5 using QDA.

Table 3 presents the results of our baseline models employing the discussed feature vectors. The best results were obtained using the Support Vector Machine classifier, the RF classifier, and the Simple Perceptron with various embedding strategies, as it achieved an accuracy of 0.95 to 0.96. The RF Classifier, Simple Perceptron, and SVM with TF-IDF, CV, and Wod2Vec performed better than their counterparts. Analyzing its classification report, we can see that fraudulent and threatening are the most difficult to categorize, with SVM achieving an accuracy of 96%. This demonstrates that the Support vector classifier handled these classes substantially better than the other baselines. Furthermore, RF using CV and W2V as features provided an accuracy of 95 and 94, respectively. QDA and GuassianNB are nearly comparable in many features, with a tiny variation; however, QDA fared very poorly with CV features, with an accuracy score of just 64%. In most basic scenarios, overlapping classes that may be used interchangeably are challenging to manage. Our key goal is to create a mechanism to handle these overlapping classes better. Figure 5 presents the confusion matrix of the two best-performing baseline models.



(a) Confusion matrix of support vector machines with tf-idf (b) Confusion matrix of Simple perceptron with tf-idf

Fig. 5. Confusion matrix of our two best-performing baseline models

4.2 Deep learning models and their results

With more deep networks, results were almost the same or marginally better as we have achieved with baselines, as shown in Table 4. At first, we train vector length of 500d FastText embeddings, and a FastText model like classifier Joulin *et al.* [37] we achieve 92.7% accuracy on the test set. We also use Bi-GRU with 300d FastText word vectors, LSTM, RNNs, and CNN. We get a 93.8%, 95.1%, 92.1%, and 94.1% accuracy.

Table 3. baseline models and their results.

Classifier	Feature	Accuracy
KNN	TF-IDF	89%
	CV	86%
	W2V	87%
SVM	TF-IDF	93%
	CV	93%
	W2V	96%
RF	TF-IDF	94%
	CV	94%
	W2V	95%
AdaBoost	TF-IDF	91%
	CV	93%
	W2V	90%
Simple Perceptron	TF-IDF	95%
	CV	94%
	W2V	96%
GaussianNB	TF-IDF	86%
	CV	75%
	W2V	78%
QDA	TF-IDF	61%
	CV	64%
	W2V	71%

Table 4. results of deep learning models on SeaFACED corpus

Model	Results
FastText	92.7%
FastText+BiGRU	93.8%
FastText+RNN	94.0%
FastText+LSTM	95.0%
FastText+CNN	92.0%

4.3 Pretrained Models & results

BERT is our top pick for a pre-trained model. BERT embeddings are fine-tuned on our dataset. Using the traditional BERT model, we successfully attained an around 97% accuracy rate. As an alternative, we employ Distil-BERT [54], a lightweight and compact variant of BERT. Distil-BERT provides a 98% success rate in our tests. In comparison to BERT, the training period is drastically reduced. Finally, we explore XLM-Roberta, a multilingual model published by FacebookAI group [19] on our dataset. The results are almost similar to the Distil-Bert model, i.e., the accuracy of 97.63%. Both models run for more epochs than BERT and perform well on the SeFACED corpus.

5 ABLATION STUDY

This section discusses changes to the BERT architecture and the effects of such changes on accuracy. In six parts, In terms of transfer learning, we give ablation research:

- BERT is used in its most basic form (Vanilla), with the whole architecture frozen and a feed-forward network with a softmax activation layer added at the end. The failure affects only the feed-forward network, as shown in Figure 6. Furthermore, the results of the ablation study are shown in Table 5. It can be seen that BERT Unfreeze + LSTM achieves the highest accuracy in comparison with other methods.

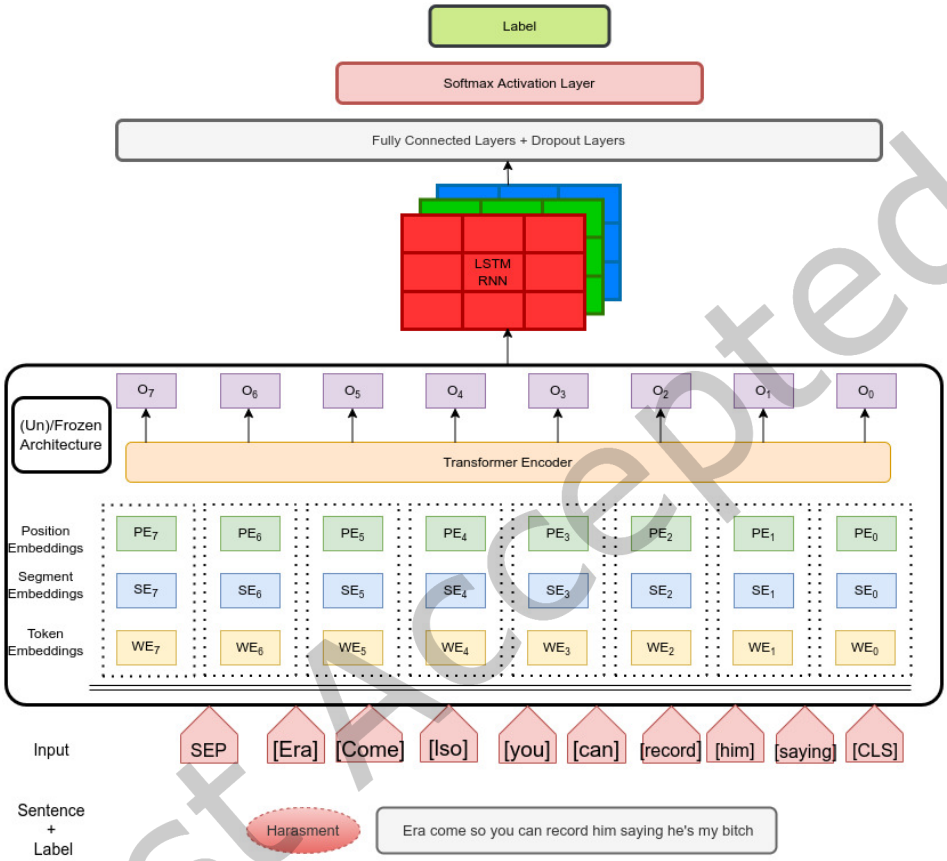


Fig. 6. BERT in its vanilla form with LSTM/RNN on top of that, by just freezing the complete architecture, i.e., using the BERT embeddings and unfreezing the complete architecture with LSTM/RNN/CNN.

- Second, we unfreeze the BERT design and replace the top layer with a softmax-activated feed-forward network. The BERT and feed-forward network topologies share the loss.
- We also freeze the BERT architecture and build convolution and dense layers on top of it. Only CNN and fully linked layers receive updates.
- We fine-tuned BERT by adding extra CNN and dense layers on top of the BERT, and the entire BERT architecture was utilized this time.
- We also investigate Seq2Seq models such as RNN and LSTM with BERT embeddings, i.e., by freezing the BERT architecture and inserting the LSTM and RNN layers at the top layer of the BERT.
- We employed RNN and LSTM with BERT architecture; that is, we used the entire BERT architecture and added layers on top of the BERT. We used LSTM, RNN, and Convolutional layers.

Table 5. Results of Ablation Study on SeaFACED corpus

Model	**Accuracy %**
BERT Freeze	90
BERT Unfreeze	94
BERT Freeze + CNN	93
BERT Unfreeze + CNN	95
BERT Freeze + LSTM	95
BERT Unfreeze + LSTM	96

6 CONCLUSION

This paper proposed a novel method for categorizing E-mails into four separate classes: Normal, Fraudulent, Threatening, and Suspicious E-mails, based on Transformer topologies that manage not just short sequences but also extensive dependencies of large characters. This approach allows for more accurate and dependable systems, with the ability to explain to the user why an email was classified as harassing, threatening, or normal. The authors also developed an evaluation tool specifically for this objective. The contributions of this study include the creation of the SeFACED corpus, which is expected to attract researchers outside of the email forensics community. This corpus provides a unique context for researchers interested in email forensics, especially those using fallacies, threats, and emotions in their approaches. Additionally, the authors tested numerous BERT-based models and found that they can learn contextual information from the text to categorize emails into sub-classes properly. The main result of this study is the achievement of the best accuracy of 98% using transformers-based architectures, such as BERT, surpassing existing studies. The findings suggest that this method can improve the investigation of spam and fraud more efficiently and effectively and eventually improve the overall experience for email users. One limitation of this study is that the corpus only includes text data, and the authors are currently working with other researchers to expand the corpus to include audio and video data. Another challenge is the continuous evolution of spam tactics, which requires regular updates to the corpus and models to maintain their effectiveness. In the future, this research will overcome the above limitations and create an approach that will help other researchers in email forensics.

REFERENCES

[1] Ahmed Abbasi, Abdul Rehman Javed, Farkhund Iqbal, Zunera Jalil, Thippa Reddy Gadekallu, and Natalia Kryvinska. 2022. Authorship identification using ensemble learning. *Scientific Reports* 12, 1 (2022), 1–16.

[2] Ahmed Abbasi, Abdul Rehman Rehman Javed, Amanullah Yasin, Zunera Jalil, Natalia Kryvinska, and Usman Tariq. 2022. A Large-Scale Benchmark Dataset for Anomaly Detection and Rare Event Classification for Audio Forensics. *IEEE Access* 10 (2022), 38885–38894.

[3] Nakul Agarwal, Vineeth N Balasubramanian, and CV Jawahar. 2018. Improving multiclass classification by deep networks using dagsvm and triplet loss. *Pattern Recognition Letters* 112 (2018), 184–190.

[4] Adnan Ahmed, Abdul Rehman Javed, Zunera Jalil, Gautam Srivastava, and Thippa Reddy Gadekallu. 2021. Privacy of web browsers: a challenge in digital forensics. In *International Conference on Genetic and Evolutionary Computing*. Springer, 493–504.

[5] Naeem Ahmed, Rashid Amin, Hamza Aldabbas, Deepika Koundal, Bader Alouffi, and Tariq Shah. 2022. Machine learning techniques for spam detection in email and IoT platforms: analysis and research challenges. *Security and Communication Networks* 2022 (2022), 1–19.

[6] Usman Ahmed, Rutvij H Jhaveri, Gautam Srivastava, and Jerry Chun-Wei Lin. 2022. Explainable deep attention active learning for sentimental analytics of mental disorder. *Transactions on Asian and Low-Resource Language Information Processing* (2022).

[7] Waqas Ahmed, Faisal Shahzad, Abdul Rehman Javed, Farkhund Iqbal, and Liaqat Ali. 2021. Whatsapp network forensics: Discovering the ip addresses of suspects. In *2021 11th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*. IEEE, 1–7.

[8] Yusra Al Balushi, Hothefa Shaker, and Basant Kumar. 2023. The Use of Machine Learning in Digital Forensics. In *1st International Conference on Innovation in Information Technology and Business (ICIITB 2022)*. Atlantis Press, 96–113.

[9] AM Al-Zoubi and H Faris. 2018. J. f. Alqatawna, and MA Hassonah,"Evolving Support Vector Machines using Whale Optimization Algorithm for spam profiles detection on online social networks in different lingual contexts,". *Knowledge-Based Systems* 153 (2018), 91–104.

[10] Abdullah Alqahtani, Habib Ullah Khan, Shtwai Alsubai, Mohemmed Sha, Ahmad Almadhor, Tayyab Iqbal, and Sidra Abbas. 2022. An efficient approach for textual data classification using deep learning. (2022).
[11] I Basyar, Murdiansyah DT Adiwijaya, and DT Murdiansyah. 2020. Email spam classification using gated recurrent unit and long short-term memory. *Journal of Computer Science* 16, 4 (2020), 559–567.
[12] Gourav Bathla and Adarsh Kumar. 2021. Opinion spam detection using Deep Learning. In *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 1160–1164.
[13] Asma Belhadi, Youcef Djenouri, Gautam Srivastava, and Jerry Chun-Wei Lin. 2023. Fast and Accurate Framework for Ontology Matching in Web of Things. *ACM Transactions on Asian and Low-Resource Language Information Processing* (2023).
[14] Diego Benito, Oscar Araque, and Carlos A. Iglesias. 2019. GSI-UPM at SemEval-2019 Task 5: Semantic Similarity and Word Embeddings for Multilingual Detection of Hate Speech Against Immigrants and Women on Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 396–403. https://doi.org/10.18653/v1/S19-2070
[15] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146. https://doi.org/10.1162/tacl_a_00051
[16] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[17] Jie Cao and Chengzhe Lai. 2020. A bilingual multi-type spam detection model based on M-BERT. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 1–6.
[18] Gurpal Singh Chhabra and Dilpreet Singh Bajwa. 2015. Review of e-mail system, security protocols and email forensics. *International Journal of Computer Science & Communication Networks* 5, 3 (2015), 201–211.
[19] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747
[20] Michael Crawford and Taghi M Khoshgoftaar. 2021. Using inductive transfer learning to improve hotel review spam detection. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 248–254.
[21] Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. 2020. Prta: A System to Support the Analysis of Propaganda Techniques in the News. *arXiv* (2020), arXiv–2005.
[22] S Jancy Sickory Daisy and A Rijuvana Begum. 2021. Smart material to build mail spam filtering technique using Naive Bayes and MRF methodologies. *Materials Today: Proceedings* 47 (2021), 446–452.
[23] Bilge Kagan Dedeturk and Bahriye Akay. 2020. Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. *Applied Soft Computing* 91 (2020), 106229.
[24] Ahmet Demirkaya, Jiasi Chen, and Samet Oymak. 2020. Exploring the role of loss functions in multiclass classification. In *2020 54th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 1–5.
[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423
[26] M Muztaba Fuad, Debzani Deb, and M Shahriar Hossain. 2004. A trainable fuzzy spam detection system. In *Proc. of the 7th Int. Conf. on Computer and Information Technology*.
[27] Surajit Giri, Siddhartha Banerjee, Kunal Bag, and Dipanjan Maiti. 2022. Comparative Study of Content-Based Phishing Email Detection Using Global Vector (GloVe) and Bidirectional Encoder Representation from Transformer (BERT) Word Embedding Models. In *2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*. IEEE, 01–06.
[28] Bahia Halawi, Azzam Mourad, Hadi Otrok, and Ernesto Damiani. 2018. Few are as good as many: An ontology-based tweet spam detection approach. *IEEE Access* 6 (2018), 63890–63904.
[29] Maryam Hina, Mohsin Ali, Abdul Rehman Javed, Fahad Ghabban, Liaqat Ali Khan, and Zunera Jalil. 2021. Sefaced: Semantic-based forensic analysis and classification of e-mail data using deep learning. *IEEE Access* 9 (2021), 98398–98411.
[30] Maryam Hina, Mohsan Ali, Abdul Rehman Javed, Gautam Srivastava, Thippa Reddy Gadekallu, and Zunera Jalil. 2021. Email Classification and Forensics Analysis using Machine Learning. In *2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI)*. IEEE, 630–635.
[31] Maryam Hina, Mohsan Ali, Abdul Rehman Javed, Gautam Srivastava, Thippa Reddy Gadekallu, and Zunera Jalil. 2021. Email Classification and Forensics Analysis using Machine Learning. In *2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI)*. IEEE, 630–635.

[32] Yaoshiang Ho and Samuel Wookey. 2019. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access* 8 (2019), 4806–4813.
[33] Gauri Jain, Manisha Sharma, and Basant Agarwal. 2019. Spam detection in social media using convolutional and long short term memory neural network. *Annals of Mathematics and Artificial Intelligence* 85, 1 (2019), 21–44.
[34] Abdul Rehman Javed, Waqas Ahmed, Mamoun Alazab, Zunera Jalil, Kashif Kifayat, and Thippa Reddy Gadekallu. 2022. A comprehensive survey on computer forensics: State-of-the-art, tools, techniques, challenges, and future directions. *IEEE Access* (2022).
[35] Abdul Rehman Javed and Zunera Jalil. 2020. Byte-level object identification for forensic investigation of digital images. In *2020 International Conference on Cyber Warfare and Security (ICCWS)*. IEEE, 1–4.
[36] Abdul Rehman Javed, Zunera Jalil, Wisha Zehra, Thippa Reddy Gadekallu, Doug Young Suh, and Md Jalil Piran. 2021. A comprehensive survey on digital video forensics: Taxonomy, challenges, and future directions. *Engineering Applications of Artificial Intelligence* 106 (2021), 104456.
[37] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 427–431. https://www.aclweb.org/anthology/E17-2068
[38] Akash Junnarkar, Siddhant Adhikari, Jainam Fagania, Priya Chimurkar, and Deepak Karia. 2021. E-mail spam classification via machine learning and natural language processing. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. IEEE, 693–699.
[39] Sanaa Kaddoura, Ganesh Chandrasekaran, Daniela Elena Popescu, and Jude Hemanth Duraisamy. 2022. A systematic literature review on spam content detection and classification. *PeerJ Computer Science* 8 (2022), e830.
[40] Hannah Kim and Young-Seob Jeong. 2019. Sentiment classification using convolutional neural networks. *Applied Sciences* 9, 11 (2019), 2347.
[41] Piotr Kłosowski. 2018. Deep learning for natural language processing and language modelling. In *2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. IEEE, 223–228.
[42] Priti Kulkarni, Jatinderkumar R Saini, and Haridas Acharya. 2020. Effect of header-based features on accuracy of classifiers for spam email classification. *International Journal of Advanced Computer Science and Applications* 11, 3 (2020).
[43] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.
[44] Yuxin Liu, Li Wang, Tengfei Shi, and Jinyan Li. 2022. Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and Bi-LSTM. *Information Systems* 103 (2022), 101865.
[45] Qin Luo, Bin Liu, Junhua Yan, and Zhongyue He. 2011. Design and implement a rule-based spam filtering system using neural network. In *2011 International Conference on Computational and Information Sciences*. IEEE, 398–401.
[46] Arun S Maiya. 2020. ktrain: A low-code library for augmented machine learning. *arXiv preprint arXiv:2004.10703* (2020).
[47] RAZA Mansoor, Nathali Dilshani Jayasinghe, and Muhana Magboul Ali Muslam. 2021. A comprehensive review on email spam classification using machine learning algorithms. In *2021 International Conference on Information Networking (ICOIN)*. IEEE, 327–332.
[48] Weilong Mo, Xiaoshu Luo, Yexiu Zhong, and Wenjie Jiang. 2019. Image recognition using convolutional neural network combined with ensemble learning algorithm. In *Journal of Physics: Conference Series*, Vol. 1237. IOP Publishing, 022026.
[49] Kamran Morovati and Sanjay S Kadam. 2019. Detection of Phishing Emails with Email Forensic Analysis and Machine Learning Techniques. *International Journal of Cyber-Security and Digital Forensics* 8, 2 (2019), 98–108.
[50] Rakesh Nayak, Salim Amirali Jiwani, and B Rajitha. 2021. Spam email detection using machine learning algorithm. *Materials Today: Proceedings* (2021).
[51] Ahmad Nsouli, Azzam Mourad, and Danielle Azar. 2018. Towards proactive social learning approach for traffic event detection based on arabic tweets. In *2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC)*. IEEE, 1501–1506.
[52] Khan Farhan Rafat, Qin Xin, Abdul Rehman Javed, Zunera Jalil, and Rana Zeeshan Ahmad. 2022. Evading obscure communication from spam emails. *Mathematical Biosciences and Engineering* 19, 2 (2022), 1926–1943.
[53] Nadjate Saidani, Kamel Adi, and Mohand Said Allili. 2020. A semantic-based classification approach for an enhanced spam detection. *Computers & Security* 94 (2020), 101716.
[54] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).
[55] Robert E. Schapire. 1999. A Brief Introduction to Boosting. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2 (IJCAI'99)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1401–1406.
[56] Alexander Semenov, Vladimir Boginski, and Eduardo L Pasiliao. 2019. Neural Networks with Multidimensional Cross-Entropy Loss Functions. In *International Conference on Computational Data and Social Networks*. Springer, 57–62.
[57] Neha Sharma, Vibhor Jain, and Anju Mishra. 2018. An analysis of convolutional neural networks for image classification. *Procedia computer science* 132 (2018), 377–384.

[58] Vishnu Dutt Sharma, Santosh Kumar Yadav, Sumit Kumar Yadav, Kamakhya Narain Singh, and Suraj Sharma. 2021. An effective approach to protect social media account from spam mail–a machine learning approach. *Materials Today: Proceedings* (2021).
[59] Jitendra Nath Shrivastava and Maringanti Hima Bindu. 2014. E-mail spam filtering using adaptive genetic algorithm. *International Journal of Intelligent Systems and Applications* 6, 2 (2014), 54–60.
[60] Amar Singh, Nidhi Chahal, Simranjit Singh, and Suneet Kumar Gupta. 2021. Spam Detection using ANN and ABC Algorithm. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 164–168.
[61] Peng Song, Chaoyang Geng, and Zhijie Li. 2019. Research on text classification based on convolutional neural network. In *2019 International Conference on Computer Network, Electronic and Automation (ICCNEA)*. IEEE, 229–232.
[62] Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075* (2015).
[63] Xiaoya Tang, Tieyun Qian, and Zhenni You. 2020. Generating behavior features for cold-start spam review detection with adversarial learning. *Information Sciences* 526 (2020), 274–288.
[64] Xin Tong, Jingya Wang, Changlin Zhang, Runzheng Wang, Zhilin Ge, Wenmao Liu, and Zhiyan Zhao. 2021. A content-based chinese spam detection method using a capsule network with long-short attention. *IEEE Sensors Journal* 21, 22 (2021), 25409–25420.
[65] Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A Fox. 2020. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200* (2020).
[66] Marco Vanetti, Elisabetta Binaghi, Elena Ferrari, Barbara Carminati, and Moreno Carullo. 2011. A system to filter unwanted messages from OSN user walls. *IEEE Transactions on Knowledge and data Engineering* 25, 2 (2011), 285–297.
[67] S Venkatraman, B Surendiran, and P Arun Raj Kumar. 2020. Spam e-mail classification for the Internet of Things environment using semantic similarity approach. *The Journal of Supercomputing* 76, 2 (2020), 756–776.
[68] Sandeep Verma, Sherali Zeadally, Satnam Kaur, and Ajay Kumar Sharma. 2021. Intelligent and Secure Clustering in Wireless Sensor Network (WSN)-Based Intelligent Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems* (2021).
[69] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6
[70] Chih-Hung Wu. 2009. Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. *Expert systems with Applications* 36, 3 (2009), 4321–4330.
[71] Qussai Yaseen et al. 2021. Spam email detection using deep learning techniques. *Procedia Computer Science* 184 (2021), 853–858.
[72] Ammara Zamir, Hikmat Ullah Khan, Waqar Mehmood, Tassawar Iqbal, and Abubakker Usman Akram. 2020. A feature-centric spam email detection model using diverse supervised machine learning algorithms. *The Electronic Library* (2020).
[73] Yangfan Zhou, Xin Wang, Mingchuan Zhang, Junlong Zhu, Ruijuan Zheng, and Qingtao Wu. 2019. MPCE: a maximum probability based cross entropy loss function for neural network classification. *IEEE Access* 7 (2019), 146331–146341.