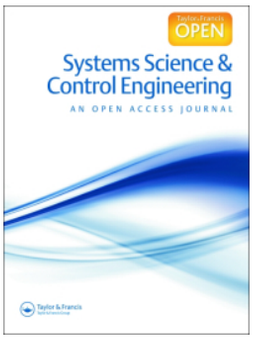

Systems Science & Control Engineering

An Open Access Journal

ISSN: 2164-2583 (Online) Journal homepage: www.tandfonline.com/journals/tssc20

Supervised methods of machine learning for email classification: a literature survey

Muath AlShaikh, Yasser Alrajeh, Sultan Alamri, Suhib Melhem & Ahmed Abu-Khadrah

To cite this article: Muath AlShaikh, Yasser Alrajeh, Sultan Alamri, Suhib Melhem & Ahmed Abu-Khadrah (2025) Supervised methods of machine learning for email classification: a literature survey, Systems Science & Control Engineering, 13:1, 2474450, DOI: 10.1080/21642583.2025.2474450

To link to this article: https://doi.org/10.1080/21642583.2025.2474450

SYSTEMS SCIENCE & CONTROL ENGINEERING: AN OPEN ACCESS JOURNAL
2025, VOL. 13, NO. 1, 2474450
https://doi.org/10.1080/21642583.2025.2474450

Taylor & Francis
Taylor & Francis Group

REVIEW ARTICLE

OPEN ACCESS Check for updates

Supervised methods of machine learning for email classification: a literature survey

Muath AlShaikh [a], Yasser Alrajeh[b], Sultan Alamri[b], Suhib Melhem[a] and Ahmed Abu-Khadrah[c]

[a]Cybersecurity Department, College of Engineering, Al Ain University, Abu Dhabi, UAE; [b]Department of Computer Science, College of Computing and Informatics, Saudi Electronic University, Riyadh, Kingdom of Saudi Arabia; [c]Department of Electrical Engineering, Al-Balqa Applied University Faculty of Engineering, As-Salt, Jordan

ABSTRACT

In today's digital landscape, email is acknowledged as a critical conduit for global data exchanges. With a surge in data volume, malefactors exploit user identities, leading to data misuse. Cybercriminals employ electronic transgressions such as phishing and spam to orchestrate security infractions. Machine learning counters these breaches using myriad techniques, demonstrating significant efficiency in identifying phishing emails. We can divide machine learning into two types: supervised and unsupervised. Supervised learning requires pre-training the model on labelled datasets, amalgamating classification, and regression learning. Notably, supervised methodologies such as support vector machines (SVMs), naive Bayes, decision trees, neural networks, random forests, and deep learning have been exploited for spam filtering. This review delves into issues concerning spam filtering and email classification through supervised machine learning techniques, offering a comprehensive evaluation of strategies, methods, performance indicators, and the benefits and drawbacks of different research. This information allows researchers to assess the efficiency and effectiveness of supervised learning algorithms, laying the foundation for advanced email categorization techniques.

ARTICLE HISTORY
Received 29 April 2024
Accepted 26 February 2025

KEYWORDS
Decision trees; naive Bayes; neural networks; phishing; random forests; support vector machine

1. Introduction

Since the advent of email, it has served as a catalyst for the development of numerous applications. As the principal medium of information exchange, the daily volume of email correspondence continues to increase. Contemporary algorithms have evolved in response to these challenges as the task of categorization is set to become increasingly onerous with the escalating complexity and dimensionality of data. Furthermore, email classification and processing present a plethora of challenges owing to their inherently dynamic, multi-topic, and perpetually evolving nature. Given that conversation topics can shift over time, discerning the most salient features of an email's myriad characteristics poses a formidable challenge.

The fast-growing use of spam-phishing techniques and unsolicited mail filters presents challenges for researchers. Therefore, security engineering professionals are paying close attention to phishing and spam email detection. Implement, design, and model spam email detection systems as necessary and modify them to consider phishing features. The fact that spammers can still easily avoid spam filtering software is another problem. By coming up with ingenious ways to get around and having their communications filtered, spammers continuously keep up with the advancement of anti-spam technology. Even though it is not simple, spammers can produce information that might get past an email filter that uses only one kind of screening. These techniques include text concealment, URL concealment, hash decryption, Bayesian sneaking, and poisoning.

One may consider a wide range of email-related elements, including email subjects, headers, payloads, and URLs. The weights assigned to each had an impact on the grouping of the ML algorithms used to produce it. A dynamic weighting technique must be utilized because each mailbox is distinct, which makes it difficult to choose and assign weights before the algorithm is executed. Machine learning is regarded as one of the most crucial methods used for email classification and spam detection. This paper reviews some of the literature on the utilization of machine learning methods in email classification and their accuracy. This literature survey will result in a comparative examination of these techniques

CONTACT Muath AlShaikh muath.alshaikh@aau.ac.ae

Supplemental data for this article can be accessed online at https://doi.org/10.1080/21642583.2025.2474450.

to identify the most efficient ones. Through this paper, other researchers will be able to expand these perspectives, discover new algorithms, and develop existing ones to avoid spam, phishing, and malware.

This paper goes through successive sections to reach its objectives. The following section discusses the email classification concept and some terminologies, such as spam, phishing, and malware. Subsequently, each supervised machine learning technique is discussed, and the literature concerns its efficiency in spam detection are reviewed. Then, a literature survey summary is introduced based on the literature reviewed in the previous section. The paper concludes with a discussion and a conclusion section showing some comparative points between these different techniques.

2. Background

Users regularly use electronic mail as an archive for documents, conversations, and other types of information, including photographs. Marketing, advertising, and newsletter mailings sent to users are becoming increasingly diversified and visual. Security systems continue to miss many spam communications, which adds more rubbish to users' mailboxes and exacerbates the overload. Emails may also have a status indicating that certain actions, such as reading or responding, are required. To-do emails are part of the domain for task/action emails (Silva, 2016).

Attackers utilize user identities and abuse these data as a result of an increase in data volume. Security breaches are made possible by cybercriminals using electronic crimes, such as phishing, spoofing, spam, threats, harassment, and terrorism. Therefore, prioritizing emails has become challenging. Because certain emails may be spam efforts, it is vital to sort them according to importance. Some threats which email users encounter are spam, phishing, and malwares.

Spam is any inappropriate and unwelcome email or message sent by a hacker to a large number of recipients. Hackers frequently utilize this approach to attract customers to utilize Internet-based services. They may send spam using numerous files attached and connections to nefarious and undesirable websites, which could result in data breaches, fraudulent transactions, and identity theft (Hu et al., 2021). Various companies develop different techniques and algorithms for efficient spam detection and filtering. Some spam filtering techniques are depicted below in Table 1.

Phishing email is a type of spam communication that is sent out to trick recipients to expose their private data, such as their password, bank account data, and debit or credit card information (Chiksa, 2022). An attacker can create a convincing and customized email by obtaining even a minimal quantity of personal information about their target. Since the majority of these phishers are able to conceal the location of their servers and operate almost anonymously, they are also difficult to stop. Phishing is an issue because its perpetrators are always coming up with inventive and new ways to trick people into thinking they are engaging with a trustworthy email or website. The second method is using filters for phishing which being designed using machine learning techniques. Moreover, training users to avoid falling for phishing scams is the third approach (Vayansky & Kumar, 2018).

Malware is a contraction of malicious software, is designed to destroy computer systems and programmes (Nikam & Deshmuh, 2022). The term 'malware' refers to a broad variety of potentially dangerous software such as ransomware, trojan horses, spyware, adware, viruses, rogue programmes, wipers, scareware, and other hazards (Akhtar & Feng, 2022). There are many ways that can be used for mitigating the impacts of the malware on computer systems. One security measure that can be used

Table 1. Spam filtering methods.

Method	Description
Standard Spam Filtering Method	Initially, artificial intelligence approaches are employed to identify spam through the use of content filters. The second stage involves implementing an email header filter, which pulls email header data out of the message. Subsequently, emails originating from the backlist file are subjected to backlist filters to weed out spam emails. Subsequently, the sender is identified using user-defined criteria and the subject line in rule-based filters. Eventually, a technique that permits the account holder to send mail is implemented, and allowance and task filters are utilized (Lee et al., 2018).
Client-Side Spam Filtering	Different guidelines and safeguards are provided by spam detection at the client point to guarantee secure communication between individuals and organizations. A client should install several different existing frameworks on his or her PC to transmit data. By composing, accepting, and handling the incoming emails, these systems establish a connection with client mail agents and filter the client's mailbox (Jain & Gupta, 2016; Pathan & Kamble, 2018).
Enterprise Level Spam Filtering	Email spam detection at the enterprise level is a technique in which various filtering frameworks are installed on the server, dealing with the mail transfer agent and classifying the collected emails into one spam or ham (Ahmed et al., 2022).
Case-Based Spam Filtering	This kind of filtering uses a collecting approach to help with many phases; in the first step, data (email) are collected. The next significant change involves the preprocessing stages of the client graphical user interface, which outlines abstraction and the choice of email data classification. The entire process is then tested using vector expression, and the data is divided into two classes: spam and legitimate email. In order to assess whether this is an email, the machine learning technique is finally extended to training sets and test sets (Ahmed et al., 2022).

to regulate and observe incoming and outgoing network traffic based on security criteria is the usage of firewalls, which come in two varieties—hardware and software. Moreover, Computer systems can be protected against malware using a variety of security software programmes, including antivirus programmes, internet security programmes, and cleanup tools (Saeed, 2020).

3. Email data processing for spam detection

The process of spam detection requires some steps which are data preprocessing, feature selection, classification techniques and evaluation metrics (Hassan & Mtetwa, 2018).

3.1. Datasets

The email dataset came from Enron and Spam Assassin, among other places. The dataset is made up of email files that are methodically arranged in directories and contain all of the email content in text format (Ghogare et al., 2024). Some information about these two types are depicted below.

- Enron Dataset: Nowadays, the Enron Corpus is the largest and one of the only publicly accessible mass collections of real emails that can be examined in the globe. Research analyzing the Enron database presented at a 2004 Conference found that the Enron corpus was suitable for Spam Classification evaluation of e-mail classification methods. Email correspondence from Enron employees, particularly top executives, and traders makes up the Enron corpus. The collection contains folder information for each of the 158 workers. The sender and recipient email addresses, the message's subject, body, text, and other email-specific technical details are all included in the folders along with the date and time of the communication.
- Spam Assassin Dataset: The Spam Assassin public corpus website is where the Spam Assassin dataset was found. It consists of two sets of data: testing and training. Every dataset consists of a sequence of plaintext emails that have been arbitrarily categorized as spam or ham. There are 9349 emails in the sample overall. The data shows that 22% of the time is spam and 78% of the time is ham. The collection contains 6951 ham emails and 2398 spam emails. The data contains the emails in plain text format. They therefore require them to transform plain text into attributes that could be associated with emails. Afterwards, they might use these attributes to run a machine learning algorithm on the emails. Initially, a range of pre-processing procedures are performed, including stemming, lemmatizing text, removing all digits, special characters, punctuation, etc., and lower-casing every character.

3.2. Data preprocessing

Data preprocessing is required prior to using a machine learning algorithm since not all email content is helpful for classifying spam. In most situations, removing particular noisy and uninformative terms can improve classifier performance and reduce the feature space dimensionality. The act of converting a particular dataset into a consistent format that is more suited for machine learning algorithms is known as data pre-processing (Mistry, 2020). The procedures required to get textual data ready for feature extraction are listed below:

- Removal of special characters: Special character such as '#', '@' etc. will be removed from the data during the preprocessing stage.
- Tokenization: To determine which terms are considered spam or ham words, the textual data is divided into individual words. Email headers, attachments, and HTML tags are removed in addition to the subject and body line text. IP addresses and domain names are also regarded as tokens.
- Stemming: This method groups together words in various inflected forms, such as plurals, gerund forms, tenses, etc.
- Case Conversion: Following tokenization of the data, words will be converted to lowercase.
- Removal of stop words: Words like 'a', 'and' and 'the' etc. are very common in English and are not helpful in identification of spam. Despite their widespread usage, these terms don't include any information that can be classified. In this step, symbols and unclear language can also be eliminated. Eliminating these terms can decrease the feature space and improve classification effectiveness.
- Representation: In this step, textual material is represented into a certain structured format that the machine learning system can comprehend.

3.3. Feature extraction

A feature selection is a technique for choosing a subset of features from data. It is often referred to as a variable selection. In machine learning, this approach is frequently used to solve problems with large dimensionality. A subset of important features is chosen, and redundant, unnecessary, and noisy features are rejected in order to streamline and summarize the data representation process (De Silva et al., 2015). Approaches to feature selection

include embedding strategies, wrapper models, and filter models.

An individual feature or quality of a phenomenon being observed might be defined as follows in the context of pattern recognition and machine learning (Wikipedia, 2017). The key text can be made available for examination through the process of transforming textual data into a particular format. TF-DIF, n-grams, bag-of-words, and other feature extraction algorithms are available for text categorization.

- Bag of words: The bag-of-words method is regarded as a well-liked and frequently applied technique for encoding textual data in documents for analysis. It is simple to comprehend and put into practice. A textual document is represented as a bag-of-words, which is essentially an unordered set of words that maintain their frequency but disregard their placement in the document.
- Term frequency – Inverse Document Frequency (TFIDF): A well-liked feature extraction method for document classification is TF-IDF. One issue with word counts and term frequencies is that frequently occurring terms tend to take centre stage in the document and may not provide as much information for the model as uncommon but domain-specific words. Rearranging the frequency of words according to their presence in each document is one method. As a result, common terms that might not have much information overall in all documents receive a lower score. The abbreviation for this document scoring method is TF-IDF, where:
  - Term Frequency (TF): The scoring of the frequency of a word in a current document.
  - Inverse Document Frequency (IDF): The scoring of the frequency of a word across all documents.

Word2vec is a method in natural language processing (NLP) for obtaining vector descriptions of words. These vectors describe information about the meaning of the word based on the surrounding words. The word2vec procedure assesses these representations by developing text in a big corpus. When trained, such a model can identify synonymous words or suggest additional words for a partial sentence. Word2vec was developed by Tomáš Mikolov and colleagues at Google and published in 2013 (Goldberg, 2014).

3.4. Classification techniques

Machine learning approaches have proven effective in the classification of spam. By using these methods, the classifier is trained with data that is extracted from labelled training datasets. Machine learning, defined as 'the ability of computers to learn without being explicitly programmed,' is a field of study that has some connection to mathematical optimization and computational statistics. Emails are categorized as either spam or ham in a binary classification issue known as spam email classification. The most widely used machine learning methods for spam categorization are Neural Network, Decision Tree, Support Vector Machine, and Naïve Bayes. When it comes to classifying spam emails, these algorithms outperform other approaches (Ahmed et al., 2022).

Table 2. Evaluation metrics.

Evaluation Metrics	Description
Accuracy score	Percentage of observations that were correctly classified.
Precision	The ability of a classifier not to label a positive sample as negative or label a negative sample as positive
Recall	The ability of a classifier to find all the positive samples in a dataset.
F1 score	The harmonic means of precision and recall.
Area Under Curve (AUC) Score	At different threshold conditions, the AUC score is an output measurement for classification problems. The degree or metric of separability is represented by the AUC. It indicates how well the model can differentiate between classes. The higher the AUC, the more accurate the model.

3.5. Evaluation metrics

The evaluation metrics given in Table 2 were used in the evaluation and comparison of the models (Mistry, 2020).

4. Supervised machine learning for email classification

Several machine learning and deep learning techniques such as Naive Bayes, Decision Trees, Neural Networks, and Random Forests have been used in email classification. Algorithms for machine learning that require marked data are included in supervised machine learning techniques. Once these models are trained using labelled training data, they may be able to predict future events. In other words, these algorithms study a training dataset that is already available before formulating a strategy to forecast success rates. The system can anticipate any new information relevant to everyone is knowledge at the time of training after receiving the requisite training (Dridi, 2022). Additionally, the learning system properly assesses the outcome of the forecast and identifies errors to improve the technique. Supervised learning was trained to forecast information using tagged data. Numerous topics, including the popularity of advertisements, classification of spam, face recognition, and object classification, can be addressed using this type of learning. In supervised

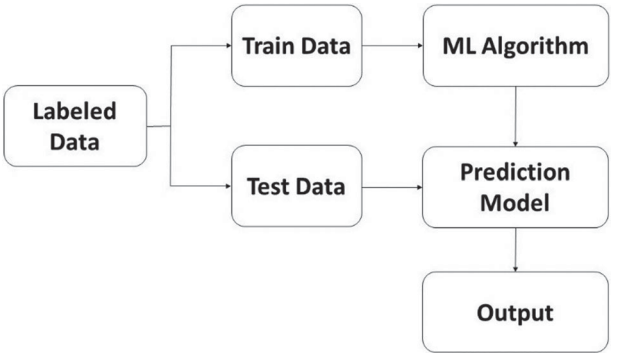

Figure 1. Process of supervised learning.

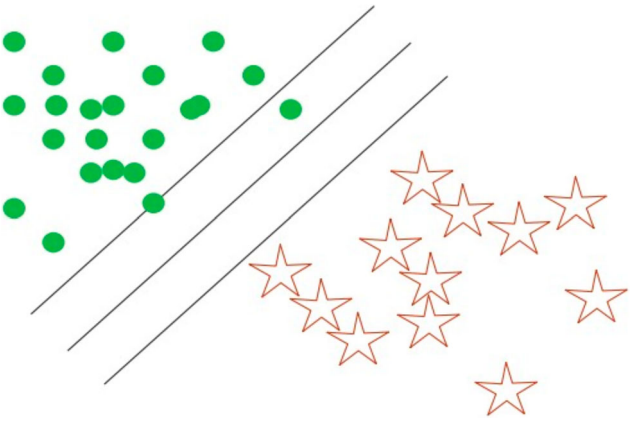

Figure 2. Support vector machine classification (Ahmed et al., 2022).

learning, regression and classification are the two primary subtypes. The supervised learning process is illustrated in Figure 1.

The following section includes a summary of the supervised learning methods and techniques, performance metrics for each supervised learning model, and an accuracy comparison of each supervised learning model.

4.1. Support Vector Machine (SVM)

A crucial and useful machine-learning paradigm is the support vector machine (SVM). The official definition of SVM is that it is a selective supervised learning classifier that creates a hyperplane to categorize fresh data and train on labelled samples. A collection of items can be classified into many classes by using decision planes. Figure 2 shows the categorization principle of the linear support vector machine (SVM). The figure contains several stars and circles, which are referred to as objects. These items can be divided into the star and dot classes. Separate lines in (e.g.) decide which things are green and brown. The brown stars at the bottom of the plane and the green dots at the top of the plane demonstrate how two identical items are divided into two different classes. Based on the training samples provided during the training phase, the model classified a new object, represented by a black circle, into one of the classes (Ahmed et al., 2022).

In a study, a support vector classifier with two Gaussian and polynomial kernels is employed in the learning stage. The results demonstrate that the proposed technique, which employs the support vector machine (SVM) algorithm and polynomial kernel, outperforms the alternatives in terms of precision, efficiency, accuracy, and F-0.969, as well as the ROC area under the curve of 0.985 (Ahmad et al., 2021).

Another study examined three machine learning classifiers, SVM, NB, and J48, and showed that SVM performed best in terms of effectiveness and false positive rate after considering a number of different factors (Shajideen & Bindu, 2018).

Another study proposed, developed, and tested a more potent SVM-based email spam detector by using a widely accepted standard database. According to the empirical simulation results, the proposed SVM-based spam detector system outperformed other previously published spam-detection algorithms when tested against the same well-known database employed in this study. The recommended SVM-based spam detector actually enhances the NSA-PSO hybrid system by 3.11%, which is currently the best among the previously disclosed spam detection techniques (Olatunji, 2019).

A multiple-learning-based solution for quick spam filtering was presented in a different paper. Laplace feature map approach was used to extract crucial characteristics from the text datasets of emails. Next, the collected features were used as inputs to the SVM spam-filtering algorithm. They conducted in-depth tests on three datasets, and the outcomes of the assessment demonstrated that the proposed method has a good accuracy-to-time ratio (Wang et al., 2021).

Another study's major goal is to categorize pertinent email header attributes, investigate them, and use support vector machines (SVM) to categorize them. The findings demonstrated that, at 88.80% and 87.20%, respectively, the classification of the email header using a Support Vector Machine (SVM) for CSDM2010 is superior to the anomaly detection challenge datasets. As a result, the SVM is a good classifier with an above-80% accuracy rate for both datasets (Khamis et al., 2020).

Machine-learning methods and their potential for spam filtering were examined in another study. The total accuracy was 91.2, 91%, 90.7%, and 88% for SVM (Support Vector Machine), NB (Gaussian naïve base), DT (Decision Tree classifier), and KNN (K-nearest neighbours), respectively (Mohey & Mohsen, 2021).

Another study used support vector machines (SVM) in conjunction with feature extraction to classify spam emails. The suggested model can analyze spam emails and distinguish between spam and non-spam emails. The proposed classifier recorded 98% accuracy (Shradhanjali & Verma, 2017).

Another study assessed the effectiveness of Nonlinear Support Vector machine (SVM)-based classifiers using two distinct kernel functions, namely, the Gaussian kernel and the linear kernel, using the SpamAssasin Public Collection dataset. Using the aforementioned dataset, the training and testing accuracies of these two kernels were evaluated in an effort to determine which kernel performs better on the dataset. Next, the classifier they tested on a few emails that were removed from Gmail's spam and inbox. For both linear and Gaussian approaches, the accuracy for testing were 98.5% and 97.1%, respectively (Singh & Pamula, 2018).

Another paper presents a latent semantic indexing-based SVM model-based email spam classification method. LSI was used as a feature-reduction technique to select features after the TF- and IDF-based features were extracted. Superior performance when compared to the current approach is guaranteed by the suggested hybrid model, which is assessed through the use of performance measurements, such as precision, recall, and f-measure (Renuka & Visalakshi, 2014).

4.2. Naive Bayes classifier (NB)

The Naive Bayes classifier is based on Bayes theorem. Given that the predictors are independent, the importance of a property influences the quantity of any other attribute. Naive Bayes classifiers do not require an iterative process, making them simple to develop and efficient enough to perform well on huge datasets. Despite its simplicity, naïve Bayes is well known for routinely outperforming other classification algorithms in a range of problems (Ahmed et al., 2022).

To improve the accuracy of the Naive Bayes Spam Filter, the authors of a study (Peng et al., 2018) suggested a new technique in their proposal to find text modifications and appropriately categorize emails as spam or ham. Using a method implemented in Python for feature extraction, which gathers semantic-based, keyword-based, and machine-learning methods the precision of Naive Bayes is boosted by over 200 percent when compared to Spam assassin. Additionally, the length of the email appeared to be related to the spam score, indicating that the controversial practice of Bayesian poisoning is actually a thing and is used by spammers (Peng et al., 2018).

Throughout the training process in another study, the authors constructed a model based on Bayesian principles. By dividing the messages into words and phrases, they examined the possibility that they would fall into spam and non-spam categories. The desired result, which has been put into practice as a web application, is the final value with maximum possibility. The accuracy of the proposed spam-filtering model was 98%. Both online and offline systems performed effectively (Chakraborty et al., 2022).

Another study analyzed various implementations of these algorithms, including SVM, NB, and K-Nearest Neighbor (KNN), in an experiment. The results indicated that NB provided the highest accuracy (Li, 2023).

The Naive Bayesian algorithm was used in another study to demonstrate spam filtering. Bigram scored better than Unigram in four different evaluation criteria, even if the accuracy rates for both word segmentation forms were higher than 0.75. A significant advantage of spam filtering may be the training of bigram word vectors, which can more accurately depict the relationship to prior words and increase accuracy (Zhu, 2023).

Another investigation confirmed that Naive Bayes classifiers performed well on datasets with a small number of emails and feature occurrences. When the dataset is created from a single email consideration, the Naive Bayes classifier can also provide the best accuracy, leading to the most unwanted emails being successfully rejected. Naive Bayes is an extremely effective and straightforward algorithm for eliminating spam emails (Sahni, 2021).

The Nave Bayes algorithm was used in another study to determine whether email content should be classified as spam. The total optimization of the conditions of the NB method considers the stochastic distribution and swarm behaviour characteristics of PSO. The Ling spam dataset was used for the purposes of this study, and its effectiveness was evaluated with regard to recall, precision, F-measure, and accuracy. The results of the evaluation show that PSO performs better than the NB method (Agarwal & Kumar, 2018).

Another study built a model for categorizing spam and legitimate emails using a Naive Bayes classification algorithm. The basis for this is what is known as the 'bag of words' or the frequency of words in emails or datasets. It computes probabilities by determining the frequencies and combinations of values in a collection of datasets. Spam emails were eliminated from the sample, the accuracy was calculated, and 93% accuracy was obtained (Tamboli & Sarode, 2021).

Another study showed that logistic regression and NB can attain 99% accuracy. The results can be utilized to build a sophisticated spam detection classifier by combining or filtering techniques (Kontsewaya et al. 2021).

Another study provided a stronger method that can recognize spam emails with over 90% accuracy compared to the spam filtering practices utilized by mail businesses today, such as Google, Yahoo, and Outlook.com. The algorithm relies on the Naïve Bayes Classification and clustering algorithm, which users can employ to generate their own dictionary files for both spam and regular emails (Cheng, 2023).

To identify spam emails, attributes were collected from the email text, and supervised machine learning algorithms were used to distinguish between them. Based on the results of the experiment, the 'Multinomial Naive Bayes' method with the feature extracted for all words without punctuation and stopwords performed the best overall, scoring 99.13% accuracy on the spam filter dataset and 98.65% accuracy on the spam detection dataset (Mondal & Dash, 2021).

Another study compared different machine-learning techniques with the suggested methodology for identifying spam emails. When evaluating models, consider a number of evaluation parameters, including accuracy, inaccuracy, evaluation time, and efficiency. KNN (96.20%), Naive Bayes (99.46%), SVM (96.90%), and rough-set classifiers (97.42%) were among the different accuracies attained by this system (Madhavan et al., 2021).

4.3. Decision tree classifier

A decision tree classifier is a group of in-depth inquiries into the characteristics of test record features. Every response they receive is followed up with a question, and they do not make a choice that is recorded until that happens (Subasi et al., 2018). Based on the given data, tree-based decision algorithms create models that are updated either regularly or iteratively. The primary goal of decision-tree-based algorithms is to predict a target variable's value, given an array of input data. This approach addresses classification and regression issues using a tree structure.

In a recent study, it is indicated that the Decision Stump algorithm was better at classifying emails and that its F-measures, precision, and recall scores were higher than those of the other comparative algorithms (Salim, 2022).

Another post suggested a keyword-based technique for identifying Arabic spam reviews. Terms in the original text that have been highlighted as crucial are known as keywords or features. The detection approach uses four different machine learning techniques: C4.5, KNN, SVM, and Naive Bayes classifiers. The findings demonstrate that, with a detection precision of 92.63%, the Decision Tree classifier outperforms other categorization techniques (Najadat et al., 2021).

Another study employed categorization based on decision trees to identify spam emails. Additionally, a novel method for selecting and reducing the characteristics is provided. The system was shown to work well, even with a limited number of characteristics and a short training dataset. Moreover, it is planned to incorporate more classifiers soon and assess their performance, in contrast to the suggested methods (Kumar, 2022).

Another study employed 'Naive Bayes, SMO, J48, and random forest' to classify spam mail using the 4601 occurrences data set. The 'random forest technique' is developed by analyzing and comparing classifiers according to their efficacy and achieves a maximum weighted F-measure of 95.50, maximum weighted accuracy, maximum weighted precision, maximum weighted recall, and maximum weighted accuracy. Compared to other methods, Naive Bayes performs better in terms of execution time (Saha et al., 2019).

A different article explained the categorization of spam emails. A total of 500 characteristics and 9324 records makes up the dataset that was employed to evaluate and train the model. This article can be helpful in eliminating unwanted commercial e-mails, malware, online fraud, and other unpleasant and difficult situations. A spam dataset was then applied after using the supervised machine-learning Random Forest method in this case. The results showed that, in comparison to other machine learning approaches, random forests offer superior accuracy (Kothapally & Kakulapati, 2021).

Another study considered a hybrid filtering mechanism with two tiers that could identify spam materials or messages. Naive Bayes and Random Forest classifiers are integrated using the majority vote method. With an average accuracy of 98.80% for CSDMC2010, 97.79% for spambase, and 98.84% for SMS spam collection datasets, the comparison findings demonstrated that the recommended model provided better results than the present conventional and contemporary algorithms and models (Juneja, 2022).

In another study, the decision tree for email categorization was constructed using the Iterative Dichotomiser 3 (ID3) algorithm and a novel feature selection method. The experimental results show that the proposed approach achieves extraordinarily high precision. The decision tree classifier had a 97.4% accuracy rate after training with 1000 emails and seven parameters. For 1500 and 3000 training examples, the precision was 97.2% and 97.32%, respectively (Čavor, 2021).

Another paper presented a system that can identify sent and received spam emails and notify relevant users. The system recognizes spam emails using a decision tree and the ID3 algorithm. For the system to identify and alert

users to new types of spam, the dataset is updated on a regular basis (Palival et al., 2018).

4.4. Artificial Neural Networks (ANN)

A computer model, known as an artificial neural network (ANN), is based on the operational features of biological neural networks, sometimes referred to as neural networks (NN). A neural network is a collection of interconnected groupings of neurons in which data are processed computationally. An ANN frequently exhibits adaptive behaviour, altering its structure in response to data circulating through the network, either from the outside or from within, during the learning period. nonlinear approaches for analyzing statistical data include neural networks. When irregular performance patterns or complex interconnections exist between inputs and outputs, they are typically used (Sutta et al., 2020). The fundamental structure of a neural network is illustrated in Figure 3.

An article examined the use of an email categorization system based on neural networks. Owing to their quick reaction and processing rates, which are byproducts of their parallel design, they are ideal for real-time applications. The system builds NN classifiers from numerous hidden layers through training. The results of the experiment indicate that the accuracy of the algorithm is clearly higher than that of the naive Bayes technique (Kumar, 2019).

Another study recommended using a CNN and a semantic graph neural network to tackle the difficult mail detection challenge. The word does not need to be included in the representation because the semantic network creates email attributes. The effectiveness of the method was evaluated using a variety of accessible databases. Tests on multiple public datasets demonstrated that the recommended approach for email classification had a high accuracy rate. Modern deep-learning-based algorithms perform better in terms of spam classification (Nisha & Muthurajkumar, 2023).

This study used an ANN Multilayer Perceptron (MLP) to categorize websites using traits associated with phishing.

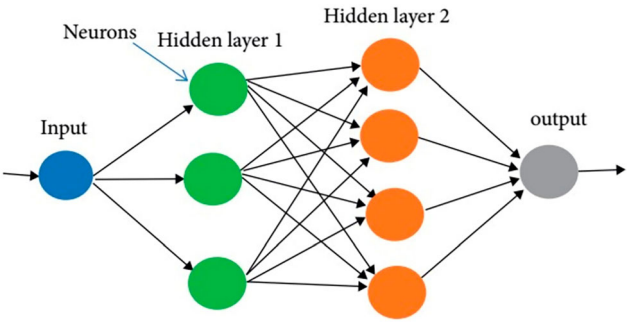

Figure 3. Basic structure of neural network (Ahmed et al., 2022).

The outcomes justify the use of an ANN-MLP to categorize websites that exhibit phishing traits. The ANN-MLP test phase showed an accuracy of 98.23% (Ferreira et al., 2018).

A technique for identifying campaigns that need to be stopped because they might be spam or phish is described in another article. It involves using real-time tracking of important campaign indications such as the open count, click count, and hard bounce count. The key feature of this system is that they do not examine the email content. Even though the proposed neural networks are quite efficient, the F1-score is higher than 0.95 for each sample used. Furthermore, this approach allows us to employ the same model regardless of the language used in the email. This technique was created and examined in cooperation with Freshmail, a significant Polish supplier of email marketing services. The excellent efficacy of the method was confirmed using the actual data provided by the company (Szpyrka et al., 2023).

A synthetic neural network model is presented for email classification (Alghoul et al., 2018). The model was trained using the feed-forward backpropagation method. The model was judged, and the result was an overall score of 85.31%. This study showed how artificial neural networks can be used to classify emails.

Emails are categorized as self or non-self, and self- and non-self-detector memory was created by reducing the redundancy from the detector set in the prior study. This is an optimized technique for email classification, as proposed in a previous study. The self- and non-self-feature vectors of the self and non-self-program are classified using a neural network classifier by means of an array of two-element self- and non-self-concentration vectors that are produced into a feature vector. The combination of neural networks and the study's earlier model will improve the false rate of their spam detector and provide both detectors with a consistent foundation for high-performance rates (Idris, 2012).

Another study assessed the effectiveness of each classifier, and the results showed that the Multilayer Perceptron classifier had the highest accuracy (98.06%). The accuracies of the naive Bayes, random forest, K Nearest neighbour, radial basis function, decision tree, artificial neural network, logistic regression, and support vector classifiers were 94.47, 95.00, 90.87, 94.02, 95.02, 98.06, 97.71, and 98%, respectively. So, ANN exhibited the best performance (Iqbal & Khan, 2022).

4.5. Summary of spam classification using machine learning

After reviewing these previous studies, Table 3 presents a summary of them, the methodology used, and the benefits and drawbacks of each study.

Table 3. Summary of the literature review on spam classification using machine learning.

Ref	Dataset	Approach	Results (%)				Merit	Demerit
			Accuracy	Precision	Recall	F-score		
Li (2023)	Kaggle	SVM, NB, KNN	99.2	-	-	-	• better data cleaning procedure • larger testing sets	• Better DL based feature learning strategies can be employed for extracting relevant features.
Hossain et al. (2019)	16,000 ham messages and 3000 spam messages	NB, SVM	90	85	88	87	• The filter is applicable to both client and server ends.	• Hard to get words and whole phrases to enter as tokens and feature vectors in the SVM.
Zhu (2023)	Kaggle	NB	79	100	20	34	• Unigram and bigram methods are used to preprocess the data.	• Time consumption of the proposed model is not included
Rusland et al. (2017)	SPAMBASE	NB	82.9	74	51	60	• The authors used two datasets	• Need to be evaluated against other methods
Shajideen and Bindu (2018)	enron1	SVM, NB, J48	94	93	90	-	• The ensemble strategy aided in obtaining a higher accuracy score	• Additional time is needed for training
Wang et al. (2021)	Enron, Genspam, PU	SVM	95.8	-	-	-	• Quick spam filtering • conducts in-depth tests on three datasets	• Need to be evaluated against other methods
Chakraborty et al. (2022)	spammassin corpus	NB	98	-	-	-	• Work on both online and offline systems	• Only a few performance metrics is evaluated to determine the model's efficiency
Murti and Naveen (2023)	Kaggle	NB, SVM, RF, DT, KNN	99.45	96.78	73.03	83.24	• Provide a detailed comparison between various ML methods	• The investigation did not measure the time consumption
Kontsewaya et al. (2021)	Kaggle	LR, NB, KNN, DT	99	97	99	98	• Provide a comparison study of various ML algorithms	• It is possible to use improved DL-based feature learning techniques
Watcharenwong and Saikaew (2017)	webcrawler	RF	98.19	98.12	-	85.12	• Combination between social and textual features gives better results	• Ignore using image features
Mani et al. (2018)	Corpus	NB, RF, SVM	87.68	89	85	–	• Using the ensemble technique made it possible to get a higher accuracy rating	• The propagation of fraudulent items in dataset adversely affect the classification process
Ban et al. (2018)	Textual data collected from Twitter and Facebook	SVM & NN	85	–	84	–	• Hybrid architecture helped to improve the classification result	• Only a few performance metrics is evaluated to determine the model's efficiency
Kumar et al. (2018)	Reviews from Yelp.com	LR, K-NN, NB, RF, SVM	76	–	–	79	• The use of both univariate and multivariate distribution across user ratings	• Efficiency needs to be enhanced
Saeed et al. (2019)	corpus	NB, SVM, K-NN, RF, NN	95.25	91.75	98.66	95.08	• Performance was increased by using Ngram feature extraction and Negation handling	• It is possible to use improved DL-based feature learning techniques

From the previous research on machine learning approaches for spam classification, as shown in Table 3, it can be inferred that researchers place a high value on machine learning techniques for their role in spam text categorization. Rule-based spam filtering solutions have limitations that machine learning can help solve. Machine learning is able to adapt to changing situations. One of the most important machine learning strategies is Support Vector Machines (SVM), a supervised learning model that examines data and finds patterns for categorization. SVMs are simple to train, and according to some researchers, they perform better than a lot of widely used techniques for classifying spam on social media. Nevertheless, with time, SVM's robustness and use for high-dimensional data decrease because of the computational complexity of the data input.

The decision tree is another machine learning technique that has been effectively used to identify spam in social media text. Decision trees (DT) require extremely minimal human input when it comes to training datasets. They have several drawbacks, including their vulnerability to overfitting of training data and the difficulty of managing tree growth without appropriate pruning. They are therefore relatively bad classifiers with limited classification accuracy.

A Naive Bayes (NB) classifier assumes that all words in the text are unrelated to one another and merely applies Bayes' theorem to the viewpoint classification of each textual data. It may be used to detect spam messages in a range of datasets with a variety of traits and properties because to its simplicity and ease of use, making it perfect for spam classification. It is also possible to enhance spam categorization tasks by employing an ensemble technique that integrates many machine learning classifiers.

Several studies on machine learning for spam categorization lead us to the conclusion that computational complexity and domain dependence are occasionally problems for ML approaches. The researchers advise using Deep Learning (DL) approaches to get around these constraints in machine learning (ML) methods for spam categorization, as certain algorithms require a lot more time and resources during training depending on the size of the dataset.

5. Deep learning methods used in spam detection

Deep learning models are becoming more and more well-liked among NLP experts since they can handle difficult issues (Kłosowski, 2018). The concept of deep learning involves training an enormous neural network with a vast quantity of data, drawing inspiration from the workings of the human brain. Automated feature extraction from the data and scalability issues can be handled by them. Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN) are the most often used deep learning models among NLP researchers.

One of the most significant and widely used Deep Learning techniques, Convolutional Neural Networks (CNN), has drawn a lot of attention lately for its ability to carry out NLP tasks. Sentiment analysis, image and text categorization, pattern recognition, and other tasks have all had success with it (Kim & Jeong, 2019; Mo et al., 2019).

Some authors employed a recurring structure for text categorization in order to extract contextual information from textual data (Lai et al., 2015). Their method surpassed CNN in text classification by extracting semantic information from text. In order to extract sequential information from textual input, the authors used the Long Short-Term Memory Network (LSTM) and developed a tree LSTM model that would work well for NLP applications (Tai et al., 2015). Another study used a Gated Recurrent Unit (GRU) model and a Long Short-Term Memory (LSTM) network to identify spam in the 34,519-record Enron e-mail spam dataset. With an accuracy of 98.39%, the LSTM model outperformed the GRU model in spam detection (Basyar et al., 2020).

The Gated Recurrent Unit-Recurrent Neural Network (GRU-RNN) was utilized in another study in order to identify spam emails from botnets (Alauthman, 2020). Their accuracy was 98.7% on the SPAMBASE dataset, which contained 2788 non-spam and 4601 spam emails. Several machine learning methods were used to assess GRU's effectiveness, however the GRU-based approach yielded the best results for spam identification. In order to determine the best features for spam email detection, some authors employed Deep Learning models like RNN in conjunction with feature selection approaches including Heatmap, Recursive Feature Elimination, and Chi-Square feature selection techniques. They got 99% accuracy on spam text data from the UCI machine learning library (Hossain et al., 2021).

Another study addressed problems with unfair representation, insufficient detection effect, and poor practicality in Chinese spam detection by utilizing a deep learning model based on LSTM and BERT (Tong et al., 2021).

Other authors employed a mix of convolution structure and bi-LSTM to extract significant and thorough semantics from a document in order to detect spam reviews pertaining to hotels. With an F1-Score of about 92.8, they might be able to exceed existing techniques in terms of classification performance (Liu et al., 2022). Numerous more studies have used Deep Learning (DL)

algorithms to detect spam, with the potential to extract textual contextual information for spam identification (Bathla & Kumar, 2021; Crawford & Khoshgoftaar, 2021).

Some authors have published analyses of deep learning methods for datasets used in spam detection. The effectiveness of such models was evaluated after they examined several deep learning-based detection strategies. Seven groups of 35 popular cyberfacts were divided and each group was evaluated. They concluded that lexical models and other traditional machine-learning techniques cannot compete with deep-learning models for intrusion and spam detection (Ferrag et al., 2020).

Artificial intelligence, particularly machine learning, is used to identify malicious emails; however, performance varies depending on the datasets utilized. An earlier study found that user feedback and actual assessment are essential, and that supervised learning may be helpful. These results motivated the implementation of an empirical study to assess the efficacy of conventional learning techniques in three email detection settings. The results of this study support earlier findings with more than 900 participants and demonstrate that LibSVM and SMO-SVM can outperform other selected algorithms (Li et al., 2022).

Another study described a novel method based on deep learning (DL) techniques in this research. In order to identify spammers, the method makes use of both the text of tweets and user meta-data (such as the age of an account, the number of followers/followers, and so forth). They demonstrate an increase in performance when utilizing their approach by comparing its performance on two separate real-world datasets with five ML-based and two DL-based state of the art approaches (Alom et al., 2020).

A different study found that the majority of routine actions and situations frequently employed deep-based approaches. The development of deep learning approaches is to blame for this. This study compared the accuracy, model strengths, and model weaknesses of the spam filtering techniques of Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and naive Bayes. The four criteria, accuracy, precision, recall, and F1-score, were used to evaluate each model using the publicly accessible dataset. It is ultimately determined that the Naive Bayes model, which can successfully handle the danger of spam, obtains the best performance among all these methods (Huang, 2023).

Another study introduced Cospam, a new approach to spam detection in IoT applications. First, the speaker and voice contents are interpreted as different timestamp feature series. Second, a cooperative neural network model was applied. The Bi-AE, GCN, and LSTM model are the three models that constitute the collaborative model. These models were used to identify the types of users. Finally, several trials were conducted to evaluate the proposed method. When compared to the techniques currently in use to identify spammers, the proposed model could attain 5% greater accuracy. Cospam takes longer than other tactics, because there are many options (Guo et al., 2020).

A novel model called the Context-LSTM-CNN model was proposed in another study because conventional classification techniques only consider the smaller context of sentences. Sentences are converted into words using word embedding, and then bi-directional LSTM is applied to these words in this model. CNN is then performed based on the outcome of the preceding phase (Ombabi et al., 2017). After reviewing these previous studies, Supplementary Table 1-Table S1 presents a summary of them, the methodology used, and the benefits and drawbacks of each study.

The foundation of various downstream activities across multiple data modalities is thought to be Pretrained Foundation Models (PFMs). Large-scale data is used to train a PFM (such as BERT [Wang et al., 2024], BERT-BKD [Sun et al., 2019], fastBERT [Liu et al., 2020], Q8BERT [Zafrir et al., 2022], WebGPT [Nakano et al., 2021], ChatGPT, GPT-4 [OpenAI, 2023]), which offers a reliable parameter initialization for a variety of downstream applications. BERT learns bidirectional encoder representations using Transformers, which are contextual language models trained on huge datasets, as opposed to previous approaches that use convolution and recurrent modules for feature extraction. Likewise, the Generative Pretrained Transformer (GPT) technique uses Transformers as feature extractors and uses an autoregressive paradigm to train on big datasets. ChatGPT has recently shown notable performance with autoregressive language models with zero-shot or few-shot prompting in large language models (Zhou et al., 2024).

Based on the above summary, the performance of the spam detection model is undoubtedly improved by these Deep Learning techniques, which also help to lessen the consequences of over-fitting that are observed in machine learning models. Deep learning approaches don't require a lot of processing power or a manual feature extraction process, in contrast to machine learning techniques. It is particularly successful at extracting spam data from text and can adapt to a wide variety of spam content prevalent in social media language. We can infer from earlier studies that the performance of spam categorization is enhanced when word-embedding techniques are combined with Deep Learning approaches. Moreover, less training data makes it harder to prevent over-fitting, and unlabelled text in the input corpus will result in worse performance. Text classification accuracy

is increased while saving a significant amount of labour and resources thanks to the deep learning approach.

6. Discussion and conclusion

The goal of this survey study was to help investigators comprehend current approaches to email classification by compiling a literature review of the email classification field through supervised machine learning methods. Review articles on supervised machine learning for email classification released after 2017 were thoroughly examined. A few logic factors were examined in the selected studies, including classification methods, performance measures, datasets, and feature sets utilized in various applications for email classification. The study also provides a summary of the merits and demerits of some of the literature reviewed. The review of existing literature in this paper essentially deals with traditional machine learning approaches for e-mail classification: decision trees, support vector machines, and k-nearest neighbours. In addition, recent developments regarding state-of-the-art approaches to NLP are mentioned: deep learning-based approaches like BERT and LSTM that show an extreme rise in performance in executing the tasks related to the classification of e-mails.

Although this review focuses primarily on studies published after 2017, some studies published before that date have been included for historical and contextual purposes. The goal is to provide a comprehensive background to the evolution of email classification techniques and to highlight the foundations upon which recent research has been built.

The study of these literature reviews indicates that the most popular techniques in supervised machine learning are SVM, decision trees, and Naive Bayes method. A review of performance metrics used in earlier research revealed that the most frequently employed metrics to assess the effectiveness of email classifiers were precision, recall, accuracy, f-measure, false positive rate, false negative rate, and error rate. SVM typically produces the best results, while Nave Bayes frequently outperforms it. However, because SVM is not compared to all other algorithms, it should not be regarded as the best method alone.

The researchers noticed that when applied to real-world data, models developed using synthetic datasets exhibit poor performance. Consequently, tests and systems should be trained on real-world data in the future rather than manually created datasets. Moreover, adopting hybrid algorithms will enable us to achieve improved accuracy and efficiency. Future feature extraction methods could benefit from deep learning.

The following list includes potential future study topics and unresolved research issues in the field of spam detection.

- Nearly all studies assessed their outcomes in terms of recall, precision, and accuracy; however, the machine learning model time complexity should be considered as an evaluation parameter.
- A spam detection and filtration system must be reliable and secure to improve the accuracy and produce consistent results.
- Each publicly accessible dataset contains only a small number of properties. Consequently, the developed models or algorithms are not as accurate as they may be, given the need for more features.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Authors contributions statement

Muath was involved in the conception and design of the concepts, Yasser prepared the analysis and interpretation of the data and the drafting of the paper, Sultan prepared the comparison, Suhib revised the comments from the reviewers and handled them; Ahmad prepared the template and references. Moreover, All authors have read and approved the final version of the manuscript.

Data availability statement

The authors agree to make data and materials supporting the results or analyses presented in our paper available upon reasonable request. The data in this paper are openly available in the paper's references.

ORCID

Muath AlShaikh http://orcid.org/0000-0001-5550-7659

References

Agarwal, K., & Kumar, T. (2018, June). Email spam detection using integrated approach of Naïve Bayes and particle swarm optimization. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 685–690). IEEE.

Ahmad, S. B. S., Rafie, M., & Ghorabie, S. M. (2021). Spam detection on Twitter using a support vector machine and users' features by identifying their interactions. *Multimedia Tools and Applications*, *80*(8), 11583–11605. doi:10.1007/s11042-020-10405-7

Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi, B., & Shah, T. (2022). Machine learning techniques for spam detection in email and IoT platforms: Analysis and research challenges. *Security and Communication Networks*, *2022*(1), 1862888. doi:10.1155/2022/1862888

Akhtar, M. S., & Feng, T. (2022). Malware analysis and detection using machine learning algorithms. *Symmetry*, *14*(11), 2304. https://doi.org/10.3390/sym14112304

Alauthman, M. (2020). Botnet spam e-mail detection using deep recurrent neural network. *International Journal of Emerging Trends in Engineering Research*, *8*(5), 1979–1986. doi:10.30534/ijeter/2020/83852020

Alghoul, A., Al Ajrami, S., Al Jarousha, G., Harb, G., & Abu-Naser, S. (2018). Email classification using artificial neural network. *International Journal of Academic Engineering Research (IJAER)*, *2*(11), 8–14.

Alom, Z., Carminati, B., & Ferrari, E. (2020). A deep learning model for Twitter spam detection. *Online Social Networks and Media*, *18*(8), 100079. https://doi.org/10.1016/j.osnem.2020.100079

Ban, X., Chen, C., Liu, S., Wang, Y., & Zhang, J. (2018, December). Deep-learnt features for Twitter spam detection. In *2018 International symposium on security and privacy in social networks and big data (SocialSec)* (pp. 208–212). IEEE.

Basyar, I., Adiwijaya, M. D., & Murdiansyah, D. (2020). Email spam classification using gated recurrent unit and long short-term memory. *Journal of Computer Science*, *16*(4), 559–567.

Bathla, G., & Kumar, A. (2021, August). Opinion spam detection using Deep Learning. In *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)* (pp. 1160–1164). IEEE.

Čavor, I. (2021, February). Decision Tree Model for Email Classification. In *2021 25th International Conference on Information Technology (IT)* (pp. 1–4). IEEE.

Chakraborty, A., Das, U. K., Sikder, J., Maimuna, M., & Sarek, K. I. (2022, October). Content based email spam classifier as a web application using naive bayes classifier. In *International Conference on Intelligent Computing & Optimization* (pp. 389–398). Springer International Publishing.

Cheng, S. (2023). Classification of spam e-mail based on naïve Bayes classification model. *Highlights in Science, Engineering and Technology*, *39*, 749–753. doi:10.54097/hset.v39i.6640

Chiksa, T. (2022, January). Phishing e-mail detection by using machine learning techniques. http://repository.smuc.edu.et/bitstream/123456789/6920/1/Tariku_Yabshe_January_2022.pdf.

Crawford, M., & Khoshgoftaar, T. M. (2021, August). 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)* (pp. 248–254). IEEE.

De Silva, A. M., Leong, P. H., De Silva, A. M., & Leong, P. H. (2015). Grammar based feature generation. *Grammar-based feature generation for time-series prediction,* 35–50.

Dridi, S. (2022). Supervised learning – A systematic literature review.

Ferrag, M., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications*, *50*, 102419. doi:10.1016/j.jisa.2019.102419

Ferreira, R., Martiniano, A., Napolitano, D., Romero, M., Gatto, D., Farias, E., & Sassi, R. (2018). Artificial neural network for websites classification with phishing characteristics. *Social Networking*, *7*(2), 97–109. doi:10.4236/sn.2018.72008

Juneja, K. (2022). Two-phase fuzzy feature-filter based hybrid model for spam classification. *Journal of King Saud University-Computer and Information Sciences*, *34*(10), 10339–10355.

Ghogare, P. P., Dawoodi, H. H., & Patil, M. P. (2024). Enhancing spam email classification using effective preprocessing strategies and optimal machine learning algorithms. *Indian Journal of Science and Technology*, *17*(15), 1545–1556. doi:10.17485/IJST/v17i15.2979

Goldberg, O. L. Y. (2014). word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arxiv. org.

Guo, Z., Shen, Y., Bashir, A., Imran, M., Kumar, N., Zhang, D., & Yu, K. (2020). Robust spammer detection using collaborative neural network in internet of thing applications. *IEEE Internet of Things Journal*, *8*(12):9549–9558. https://doi.org/10.1109/JIOT.2020.3003802

Hassan, M. A., & Mtetwa, N. (2018, November). Feature extraction and classification of spam emails. In *2018 5th international conference on soft computing & machine intelligence (ISCMI)* (pp. 93–98). IEEE.

Hossain, F., Uddin, M. N., & Halder, R. K. (2021, April). Analysis of optimized machine learning and deep learning techniques for spam detection. In *2021 IEEE international IOT, electronics and mechatronics conference (IEMTRONICS)* (pp. 1–7). IEEE.

Hossain, M. S., Zubair, M., Rahman, M. O., Patwary, M. K. H., & Rajib, M. G. S. (2019, May). A modified Naïve Bayesian-based spam filter using support vector machine. In *2019 1st international conference on advances in science, engineering and robotics technology (ICASERT)* (pp. 1–7). IEEE.

Hu, Y., Zhang, Y., & Gong, D. (2021). Multiobjective particle swarm optimization for feature selection with fuzzy cost. *IEEE Transactions on Cybernetics*, *51*(2), 874–888. doi:10.1109/TCYB.2020.3015756

Huang, Y. (2023). Spam classification based on machine learning algorithm. *Highlights in Science, Engineering and Technology*, *34*, 32–38. doi:10.54097/hset.v34i.5371

Idris, I. (2012). E-mail spam classification with artificial neural network and negative selection algorithm. *International Journal of Computer Science & Communication Networks*, *1*(3), 227–231.

Iqbal, K., & Khan, M. S. (2022). Email classification analysis using machine learning techniques. *Applied Computing and Informatics*.

Jain, A., & Gupta, B. (2016). A novel approach to protect against phishing attacks at client side using auto-updated white-list. *EURASIP Journal on Information Security*, *2016*(1), 9. doi:10.1186/s13635-016-0034-3

Khamis, S. A., Foozy, C. F. M., Aziz, M. F. A., & Rahim, N. (2020). Header based email spam detection framework using Support Vector Machine (SVM) Technique. In *Recent advances on soft computing and data mining: Proceedings of the fourth international conference on soft computing and data mining (SCDM 2020)*, Melaka, Malaysia, January 22–23, 2020 (pp. 57–65). Springer International Publishing.

Kim, H., & Jeong, Y. (2019). Sentiment classification using convolutional neural networks. *Applied Sciences*, *9*(11), 2347. https://doi.org/10.3390/app9112347

Kłosowski, P. (2018, September). Deep learning for natural language processing and language modelling. In 2018 signal processing: Algorithms, architectures, arrangements, and applications (SPA) (pp. 223–228). IEEE.

Kontsewaya, Y., Antonov, E., & Artamonov, A. (2021). Evaluating the effectiveness of machine learning methods for spam detection. *Procedia Computer Science*, *190*(3), 479–486. doi:10.1016/j.procs.2021.06.056

Kothapally, N., & Kakulapati, V. (2021). Classification of spam messages using random forest algorithm. *Journal of Xidian University*, *15*(8), 495–505.

Kumar, D. (2019). Email classification using artificial neural network. https://www.researchgate.net/publication/330675179_EMAIL_CLASSIFICATION_USING_ARTIFICIAL_NEURAL_NETWORK

Kumar, N. (2022). Model of decision tree for email classification. *International Journal of Science and Research (IJSR)*, *11*(7), 1502–1505. doi:10.21275/SR22722110223

Kumar, N., Venugopal, D., Qiu, L., & Kumar, S. (2018). Detecting review manipulation on online platforms with hierarchical supervised learning. *Journal of Management Information Systems*, *35*(2), 350–380. doi:10.1080/07421222.2018.1440758

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015, February). Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 29, No. 1).

Lee, D., Lee, M., & Kim, B. (2018). Deviation-based spamfiltering method via stochastic approach. *EPL (Europhysics Letters)*, *121*(6), 68004. https://doi.org/10.1209/0295-5075/121/68004

Li, Y. (2023). Algorithm comparison on email spam filtering task. *Highlights in Science, Engineering and Technology*, *34*, 143–148. doi:10.54097/hset.v34i.5436

Li, W., Ke, L., Meng, W., & Han, J. (2022). An empirical study of supervised email classification in Internet of Things: Practical performance and key influencing factors. *International Journal of Intelligent Systems*, *37*(1), 287–304. doi:10.1002/int.22625

Liu, Y., Wang, L., Shi, T., & Li, J. (2022). Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and Bi-LSTM. *Information Systems*, *103*(2), 101865. https://doi.org/10.1016/j.is.2021.101865

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Deng, H., & Ju, Q. (2020). Fastbert: A self-distilling BERT with adaptive inference time arxiv.

Madhavan, M., Pande, S., Umekar, P., Mahore, T., & Kalyankar, D. (2021). Comparative analysis of detection of email spam with the aid of machine learning approaches. *IOP Conference Series: Materials Science and Engineering*, *1022*(012113), 1–12.

Mani, S., Kumari, S., Jain, A., & Kumar, P. (2018). Spam review detection using ensemble machine learning. In Machine learning and data mining in pattern recognition: 14th international conference, MLDM 2018, New York, NY, USA, July 15-19, 2018, Proceedings, Part II 14 (pp. 198–209). Springer International Publishing.

Mistry, N. (2020). *Detecting spam in emails using advanced machine learning methods, Sudbury, Ontario, Canada* [Master's Thesis].

Mo, W., Luo, X., Zhong, Y., & Jiang, W. (2019). Image recognition using convolutional neural network combined with ensemble learning algorithm. *Journal of Physics: Conference Series*, *1237*(2), 022026. https://doi.org/10.1088/1742-6596/1237/2/022026

Mohey, H., & Mohsen, S. (2021). Using machine learning techniques for predicting email spam. *International Journal of Instructional Technology and Educational Studies*, *2*(4), 19–23.

Mondal, A., & Dash, A. (2021). Text based spam email classification using supervised machine learning. *Journal of Huazhong University of Science and Technology*, *50*(4), 1–14.

Murti, Y. S., & Naveen, P. (2023). Machine learning algorithms for phishing email detection. *Journal of Logistics, Informatics and Service Science*, *10*(2), 249–261. https://doi.org/10.33168/JLISS.2023.0217

Najadat, H., Alzubaidi, M., & Qarqaz, I. (2021). Detecting Arabic spam reviews in social networks based on classification algorithms. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *21*(1), 1–13. doi:10.1145/3476115

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., & Jiang, X. (2021). Webgpt: Browser-assisted question-answering with human feedback. *arXiv:2112.09332*.

Nikam, U. V., & Deshmuh, V. M. (2022, April). Performance evaluation of machine learning classifiers in malware detection. In 2022 IEEE international conference on distributed computing and electrical circuits and electronics (ICDCECE) (pp. 1–5). IEEE.

Nisha, S. R., & Muthurajkumar, S. (2023). Semantic graph based convolutional neural network for spam e-mail classification in cybercrime applications. *International Journal of Computers Communications & Control*, *18*(1), 1–12. https://doi.org/10.15837/ijccc.2023.1.4478

Olatunji, S. (2019). Improved email spam detection model based on support vector machines. *Neural Computing & Applications*, *31*(3), 691–699. doi:10.1007/s00521-017-3100-y

Ombabi, A. H., Lazzez, O., Ouarda, W., & Alimi, A. M. (2017, November). Deep learning framework based on Word2Vec and CNNfor users interests classification. In 2017 Sudan conference on computer science and information technology (SCCSIT) (pp. 1–7). IEEE.

OpenAI. (2023). Gpt-4 technical report.

Palival, D., Printer, K., Devre, R., & Lemos, N. (2018). Email spam filtering using decision tree algorithm. *International Journal of Scientific & Engineering Research*, *9*(3), 40–42.

Pathan, M. F. N. K., & Kamble, V. (2018). A review various techniques for content based spam filtering. *Engineering and Technology*, *4*.

Peng, W., Huang, L., Jia, J., & Ingram, E. (2018). Enhancing the naive Bayes spam filter through intelligent text modification detection. In *Proceedings of the 2018 17th IEEE international conference on trust, security and privacy in computing and communications/12th IEEE international conference on Big Data science and engineering (TrustCom/BigDataSE), New York, NY, USA*.

Renuka, K.D., & Visalakshi, P. (2014). Latent semantic indexing based SVM model for email spam classification. *Journal of Scientific & Industrial Research*, *73*, 437–442.

Rusland, N., Wahid, N., Kasim, S., & Hafit, H. (2017). Analysis of Na¨ıve Bayes algorithm for email spam filtering across multiple datasets. *IOP Conf. Series: Materials Science and Engineering*, *226*, 012091. https://doi.org/10.1088/1757-899X/226/1/012091

Saeed, M. A. H. (2020). Malware in computer systems: Problems and solutions. *IJID International Journal on Informatics for Development*, *9*(1), 1–8. https://doi.org/10.14421/ijid.2020.09101

Saeed, R., Rady, S., & Gharib, T. (2019). An ensemble approach for spam detection in Arabic opinion texts. *Journal of King Saud University - Computer and Information Sciences*, *34*(2), 1407–1416. https://doi.org/10.1016/j.jksuci.2019.10.002

Saha, S., DasGupta, S., & Das, S. K. (2019). Spam mail detection using data mining: A comparative analysis. In Smart intelligent computing and applications: proceedings of the

second international conference on SCI 2018, Volume 1 (pp. 571–580). Springer Singapore.

Sahni, R. (2021). Analysis of naıve Bayes algorithm for email spam filtering. *International Journal for Modern Trends in Science and Technology*, *7*(1), 5–9. doi:10.46501/IJMTST0701002

Salim, M. (2022). Using decision tree algorithms in detecting spam emails written in Malay: A comparison study. In *ITM web of conferences* (Vol. 42).

Shajideen, N. M., & Bindu, V. (2018, March). Spam filtering: A comparison between different machine learning classifiers. In 2018 second international conference on electronics, communication and aerospace technology (ICECA) (pp. 1919–1922). IEEE.

Shradhanjali, & Verma, T. (2017). E-mail spam detection and classification using SVM and feature extraction. *International Journal of Advance Research, Ideas and Innovations in Technology*, *3*(3), 1491–1495.

Silva, A. (2016). *Email classification: A case study*. University of Porto: Faculty of Engineering.

Singh, M., & Pamula, R. (2018, September). Email spam classification by support vector machine. In 2018 international conference on computing, power and communication technologies (GUCON) (pp. 878–882). IEEE.

Subasi, A., Alzahrani, S., Aljuhani, A., & Aljedani, M. (2018). Comparison of decision tree algorithms for spam E-mail filtering. In *Proceedings of the 2018 1st International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia*.

Sun, S., Cheng, Y., Gan, Z., & Liu, J. (2019). Patient knowledge distillation for BERT model compression. arxiv.org.

Sutta, N., Liu, Z., & Zhang, X. (2020). A study of machine learning algorithms on email spam classification. In *Proceedings of the 35th international conference, ISC high performance 2020, Frankfurt, Germany*.

Szpyrka, M., Suszalski, P., Obara, S., & Nalepa, G. (2023). Email campaign evaluation based on user and mail server response. *Applied Science*, *13*(3), 1630. https://doi.org/10.3390/app13031630

Tamboli, H., & Sarode, S. (2021). An effective spam and ham word classification using naïve Bayes classifier. *International Journal of Scientific and Research Publications*, *11*(7), 78–83. doi:10.29322/IJSRP.11.07.2021.p11510

Tai, K., Socher, R., & Manning, C. (2015). Improved semantic representations from tree-structured long short-term memory networks. http://arxiv.org/abs/1503.00075

Tong, X., Wang, J., Zhang, C., Wang, R., Ge, Z., Liu, W., & Zhao, Z. (2021). A content-based Chinese spam detection method using a capsule network with long-short attention. *IEEE Sensors Journal*, *21*(22), 25409–25420. doi:10.1109/JSEN.2021.3092728

Vayansky, I., & Kumar, S. (2018). Phishing – challenges and solutions. *Computer Fraud & Security*, *2018*(1), 15–20. doi:10.1016/S1361-3723(18)30007-1

Wang, Y., Sun, Y., Fu, Y., Zhu, D., & Tian, Z. (2024). Spectrum-BERT: Pre-training of deep bidirectional transformers for spectral classification of Chinese liquors. IEEE Transactions on Instrumentation and Measurement.

Watcharenwong, N., & Saikaew, K. (2017, July). Spam detection for closed Facebook groups. In 2017 14th international joint conference on computer science and software engineering (JCSSE) (pp. 1–6). IEEE.

Wikipedia. (2017). Machine Learning. https://en.wikipedia.org/wiki/Machine_learning

Zafrir, O., Boudoukh, G., Izsak, P., & Wasserblat, M. (2022). Q8BERT: Quantized 8bit BERT. EMC2@NeurIPS.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., … & Sun, L. (2024). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. International Journal of Machine Learning and Cybernetics, 1–65.

Zhu, Y. (2023). Naive Bayesian spam filtering. *Highlights in Science, Engineering and Technology*, *38*, 64–69. doi:10.54097/hset.v38i.5734