



RESEARCH ARTICLE

AN EMPIRICAL ANALYSIS OF FEATURE REPRESENTATION TECHNIQUES FOR EMAIL SPAM DETECTION

*Maria Qamar

Research Scholar, Department of Computer Science, Radha Govind University, Ramgarh, Jharkhand India

ARTICLE INFO

Article History:

Received 29th June, 2023
Received in revised form
18th July, 2023
Accepted 15th August, 2023
Published online 27th September, 2023

Key words:

Heat, Expansion, Motion, Temperature

*Corresponding author: Maria Qamar

ABSTRACT

Email spam has turned out to be one of the thorns in the flesh as it has been estimated that almost fifty percent of all emails are spam and it creates immense financial and security threats to individuals and organizations. This paper provides an empirical study of feature representation methods used in emails spam detection, which addresses the effect of the techniques on the performance of the classifier. We contrast the old bag-of-words and TF-IDF representations with the recent word embedding and deep contextual embedding techniques. A thorough experimental analysis is done with benchmark email corpora to determine accuracy, precision, recall, and other measures of a group of classifiers. Evidence shows that more sophisticated feature representations (word embeddings and transformer-based contextual vectors) raise spam detection accuracy by a significant margin compared to classical representations. Specifically, models that are trained on the features of deep learning have higher performance, and in our experiments, they reach the accuracy of about 99 percent. We talk about feature choice and its implications on model effectiveness and computing cost, and the issue of concept drift in spam content. Important Keywords Email spam, feature representation, machine learning, word embedding, spam detection.

Copyright©2023, Maria Qamar. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Citation: Maria Qamar. 2023. "An Empirical Analysis of Feature Representation Techniques for Email Spam Detection". *International Journal of Current Research*, 14, (09), 25789-25796.

INTRODUCTION

Unsolicited bulk email better known as spam has become a small nuisance to become a large-scale threat to the communication systems. According to recent statistics, the amount of spam email sent all over the world is about 45-50 percent of all email data. Such a large number of spam messages do not only overload the inbox of users but also cause enormous losses of money; billions of dollars are lost over the years by businesses and individuals through phishing, malware attacks, and time wasted (Karim *et al.*, 2019; Asliyukse *et al.*, 2025). The necessity of an efficient filtering is also supported by the fact that previously spam content consisted of just ads, but now it can be represented by advanced phishing messages and malware attacks (Dada *et al.*, 2019). Initial filtering methods (e.g. blacklists, regular expressions) soon became ineffective against the evolving strategies of spammers, and machine learning-based methods of automated spam email detection were developed (Dada *et al.*, 2019). Spam filters using machine learning consider the problem of detecting spam in an email as the problem of binary text classification: an email is identified as spam or ham (legitimate) depending on the properties of the email. The crucial point of such systems is the representation of features i.e. the way email data (mainly text content, but also metadata) is converted into inputs of a learning algorithm. In the early 2000s, applied bag-of-words models and Naive Bayes classifiers were pioneered on spam filtering, and it was shown that statistical text features may be useful to differentiate between spam and legitimate mail (Karim *et al.*, 2019).

As time passed by, more advanced representations were investigated, such as TF-IDF weighted term vectors and n-grams, which represent the importance of the term and word sequence, respectively (Dada *et al.*, 2019; Rapacz *et al.*, 2021). These conventional methods of feature extraction with such algorithms as Support Vector Machines (SVM) and Random Forests demonstrated rather decent accuracy in controlled experiments - typically over 90% (AbdulNabi and Yaseen, 2021; Rapacz *et al.*, 2021). Nevertheless, they lack the ability to understand the email text in a deeper semantic meaning and context. As an example, plain term-frequency models can also mark a message containing encoded text or highly complex language, which are now widespread in recent spam (Bacanin *et al.*, 2022). Over the past 10 years, natural language processing (NLP) has developed more text representations that can be used to detect spam. The application of distributed word embeddings has been one of the avenues (Bahgat *et al.*, 2018). Word2Vec and GloVe are techniques that produce continuous vectors of words that express semantic similarity, thus allowing the classifier to identify subtle relationships among words and synonyms commonly used by spammers. Later, contextual embeddings of deep transformer models such as BERT have been used to detect spam, where the model can comprehend the contextual meaning of words (Nasreen *et al.*, 2024; Boussaad *et al.*, 2025). These characterizations can distinguish between innocent and malicious application of similar vocabulary and pick up upon the syntactic and semantic subtleties. It has been demonstrated that deep contextual feature spam filters massively outperform simple spam filters based on either a static embedding or bag-of-words, particularly on

sophisticated spam like phishing emails (Zouak *et al.*, 2025; Kshirsagar *et al.*, 2025). The integration of non-textual and structural features is also another important development. The present email clients and research prototypes contain such features as the availability of the HTML code, quantity of links or images, sender reputation, and stylometric signs (Nuha and Lin, 2025). As an illustration, spam e-mails tend to have abnormal characteristics like excessive punctuations, all capital phrases, or suspicious URLs. Incorporation of such attributes with text content would increase the performance of detection (Nuha and Lin, 2025).

The emergence of the hybrid constructions - content based features and header/IP features or even network level signatures - is an indication of the arms race between the spammers and the defenders. In spite of these developments, there are still issues. Spammers keep evolving and creating countermeasures to filters like adding random noise, text pictures, or fake author styles. Such a continuous change may lead to concept drift, worsening the performance of the models that are not updated regularly (Asliyukse *et al.*, 2025). Therefore, it is high time to conduct an empirical study of the methods of feature representation to determine which of the methods is the most resilient and can be the most accurate in the conditions of the reality. We expect to comparatively analyze various methods of feature representation, both traditional and the most recent, in this research paper to the email spam detection problem.

To determine the efficacy of various learning algorithms, we do experimental designs so as to test the effect of each representation. We also go through the literature to bring our findings into the general development of the spam filtering studies. The final aim is to offer recommendations on which feature representations will be effective to improve spam detection performance and to be conscious of the quality of the performance and the flexibility. The rest of this paper is structured in the following way. Literature Survey section discusses related literature pointing out what has already been done in terms of feature extraction and selection in spam filtering. The next section, called the Research Methodology, describes our experimental framework, such as the description of the datasets, feature extraction mechanisms, and classification models. We then give the Results and Discussion and compare the performance metrics of the feature representation methods and discuss the implications. Lastly, the paper wraps up by giving a Closure on the main findings and recommendations on the future research areas.

LITERATURE SURVEY

A significant amount of literature has covered email spam detection using the machine learning framework, generating a wide variety of results which vary primarily due to the manner in which they encode email features and exploit them to generate spam detection. We divide the literature into three general categories as follows: (1) the traditional feature-based methods of machine learning, (2) the methods of feature selection and extraction based on optimization and hybridization, and (3) the methods of feature selection and extraction based on deep learning and embeddings. Older ML Solutions: The older machine learning methods of spam filtering were based on comparatively simple textual characteristics. The bag-of-words (BOW) model used along with the Naive Bayes (NB) classifier was shown to be effective, set up a baseline with which content-based filters would be evaluated (Androutopoulos *et al.*, 2000). Representing an email using a bag-of-words representation, an email is transformed into a vector of frequencies of token occurrences, and it may be normalized. This method was rather effective even though it is so simple, since many words (e.g., free, win, credit) occur in spam mail more often, NB or some other classifiers can be trained to learn a probabilistic model of spam vs. ham. Later studies enhanced the BOW by including weighting schemes such as TF-IDF (term frequency-inverse document frequency) which allow to remove the influence of common non-discriminative words. TF-IDF features were then used to apply classifiers (SVM, logistic regression and decision trees) that produced better precision and recall than basic

counts of terms (Dada *et al.*, 2019). In one case, AbdulNabi and Yaseen (2021) employed TF-IDF along with an ensemble of SVM and neural network classifiers, where the reported accuracy of detection is approximately 95-98% on benchmark corpora. Equally, the traditional models are improved by adding the n-gram (sets of n words or characters) features to reflect the patterns in phrases that are frequently observed in spam (e.g., act now or risk-free). These traditional methods are effective and understandable, but they can be easily confused by ingenious obfuscation methods employed by spammers (adding symbols in words or synonyms) as BOW and TF-IDF are not able to get the semantic association between words. In addition, they also generate feature vectors that are extremely high dimensional (one dimension per word or n-gram) and this can be problematic to generalization in case the training data is small and/or skewed. However, as recently as 2021, research still indicated that using classical features and algorithms such as Random Forest or SVM, the strong performance with an accuracy of more than 90% is possible (Rapacz *et al.*, 2021). Rapacz *et al.* (2021) offered another approach to quick classifier selection as well, and it should be stressed that, even in the old paradigm, the selection of the correct classifier and a feature subset (such as what words to select as being most informative) may significantly influence the effectiveness of spam filtering. Feature Selection and Hybrid Techniques: With the increase in email data and the variety of email data, scholars identified the significance of feature selection and dimensionality in detecting spam. Most mails have extraneous or noisy tokens (e.g. random strings, email addresses, disclaimers) that may confuse the classifiers. Such techniques as information gain, chi-square filtering, and principal component analysis (PCA) were utilized to eliminate irrelevant features of BOW/TF-IDF representations (Karim *et al.*, 2020). Bio-inspired optimization algorithms have been used in recent years to automatically pick or weight features to use in spam classification. As an example, Nasreen *et al.* (2024) proposed a new feature selection scheme with the help of Grey Wolf Optimizer (GWO) with BERT embeddings. They claimed their model (GWO-BERT) removes noisy features in text embeddings and achieved an accuracy of approximately 99.7 per cent on a large email dataset, which was much higher than the baseline models that do not optimize features.

On the same note, Abualhaj *et al.* (2025) suggested a framework Firefly Algorithm (FOA) based feature selection framework to detect spam URLs (that is, features of email hyperlinks). They identified 31 highly predictive features out of the original 72 URL features and improved the top accuracy of 99.76% with a lightweight KNN classifier, which showed that feature selection can be strategically selected to improve performance and efficiency. It has been also tested by other metaheuristic algorithms such as Genetic Algorithms, Particle Swarm Optimization, and Ant Colony Optimization to optimize feature sets or parameters of classifiers in email filtering (Bacanin *et al.*, 2022; Ruano-Ordas *et al.*, 2018). In a hybrid system, Bacanin *et al.* (2022) integrated an improved sine-cosine swarm intelligent algorithm, along with machine learning models, which optimize both the training of features subsets and classifiers. This approach was found to be more precise and recalling than the usual machine learning models on various spam benchmark datasets (Bacanin *et al.*, 2022). The optimization-oriented methods imply that in addition to the raw representation of text, it is important to choose what features (or transformations of features) to utilize. They frequently result in hybrid approaches - such as a logistic regression with reduced number of metaheuristically-selected features - which can be more effective than a pure classifier or a pure optimizer strategy on its own. The other hybrid approach to literature is the one which brings the content features and other modalities together. Other works combine network-level or header (such as IP address blacklists, sender authentication outcomes, etc.) with textual content (to enhance robustness) features (Mun *et al.*, 2025). Mun *et al.* (2025) indicate the existence of a system, known as the PhiShield, which combines AI-based content classification with third-party intelligence (like trusted malicious senders) to tailor spam detection. These systems handle situations that the pure text analysis may overlook - e.g. a spam email with harmless text, but where the spam server is under known spam domains, can be identified by analyzing the headers. Nonetheless,

their main essence is content-based, as the essence of the majority of the modern filters is supplemented with these additions to achieve defense-in-depth. Semi-supervised and unsupervised learning is also of interest in order to use unlabeled email data. Spam prototyping or grouping into clusters has been done by clustering. Karim *et al.* (2020) came up with an unsupervised framework that successfully clusters emails into spam/ham by learning distributional patterns and can be used as a preprocessing step to filtering. Alrammahi *et al.* (2025) also performed unsupervised clustering (K-means and Fuzzy C-means) on mixed email data sets, and the cluster labels were also the extra features to a supervised classifier. This method revealed underlying patterns of email information and enhanced classification performance when compared to the case of utilizing only raw features (Alrammahi *et al.*, 2025). Enhancements based on clustering are consistent with the results of Janez-Martino *et al.* (2023), who employed hierarchical clustering to cluster emails and a topic modeling method, which achieved higher spam vs. legitimate separations. The hybrid methods emphasize that a one-size-fits-all property of a feature representation is not necessarily ideal; it can be better to use a combination of different representation methods or to stack unsupervised and supervised processes. Deep Learning and Embeddings: The most recent generation of spam detectors builds upon deep neural networks and state-of-the-art NLP embeddings to automatically learn high-level features on emails. Deep learning models can learn feature hierarchies and take raw text (after simple preprocessing) and learn complex features such as word counts or ratios, rather than having to manually design them.

One of the examples is word embeddings. Bahgat *et al.* (2018) presented a variant of hybrid model whereby pretrained word embeddings (which reflect semantic similarity among the words) enter into classifier ensemble. They used their system of SVM and neural network classifiers, which worked with the embedding features, and it was more accurate than the traditional TF-IDF on the same data (Bahgat *et al.*, 2018). The effectiveness of word embeddings opened the door to even more effective contextual models. Recurrent neural networks (RNNs) such as LSTM and GRU have been used to detect spam, where it is possible to track sequences of words and long-term relationships in an email (Zavrak and Yilmaz, 2023). Zavrak and Yilmaz (2023) introduced a hierarchical hybrid deep model which combines convolutional neural networks with gated recursive units to identify email spam. Their model, which employs CNN to extract local patterns (phrases) and the attention-equipped GRU to extract context, generated an accuracy of greater than 99 percent in a variety of datasets (Zavrak and Yilmaz, 2023). Similarly, Boussaad *et al.* (2025) made use of a Sentence-BERT embedding, which has an attention-enhanced GRU and an SVM output layer and achieved an astounding 99.9 percent accuracy in one benchmark. The findings explain the fact that deep networks are capable of learning discriminative spam features (i.e., specific keyword patterns, writing styles or even implied context) that simple models cannot detect. Spam filtering has been further enhanced with the introduction of models in NLP that use transformers. Transformers such as BERT, RoBERTa and DistilBERT offer contextual word embeddings that indicate the meaning of words based on the context of surrounding text. The researchers have used finetuned transformers on spam data sets with great results. In one example, Asliyukse *et al.* (2025) said that a DistilBERT-based classifier with a small set of simple statistical features (such as email length and number of punctuation marks) had a 99.62% accuracy on a large combined email corpus - again, by far out-of-classical models. A different study by Zouak *et al.* (2025) proposed a hybrid approach, the BERT-GraphSAGE, which does not only process textual data with the help of BERT but instead also uses GraphSAGE to introduce the relationships (the emails and terms are treated as a graphical structure). This method was especially strong, with a rate of more than 99 percent on the Ling-Spam database and with an enhanced generalization to more recent spam. Transformers are good at dealing with subtle patterns in language - e.g. they are capable of detecting that a harmless-looking sentence has hidden malicious intent through subtle word usage or context that may confuse a bag-of-words model. Kshirsagar *et al.* (2025) introduced a meta-learning architecture which

combines a variety of classifiers (both classical and deep) and discovered that a variety of word vectorization schemes (including GloVe embeddings and BERT) enabled the meta-learner to achieve a higher accuracy in comparison to any single model, reaching approximately 99.05% on a hybrid dataset. Notably, such complex models are more expensive in terms of the level of computation, and they demand more data to be trained. Nevertheless, they have shown to be highly successful when used prudently (e.g., in fine-tuning an already trained BERT on a small spam dataset, as opposed to training it), in which case they are highly effective. In brief, it can be concluded that the literature demonstrates a definite trend of the spam detection improvement with the development of the feature representation. The use of simple text representations developed an early use of text which has formed the basis of the traditional methods but is competitive in limited environments because of efficiency. These classical models have been improved with optimization and hybrid techniques that have been used to select superior features and fuse complementary sources of information. Lastly, the use of deep learning and embeddings have taken the performance to a new level, with a number of more recent works demonstrating a near-perfect accuracy on standard data. Some of these reported levels of performance in the various feature techniques will be reflected in Table 1 (in the Results section). It should be mentioned that as the accuracy increases, researchers also allege the significance of generalizability - models should not only be able to overfit on known spam, but also identify new spam campaigns. Such methods as lifelong learning, training models periodically, and managing concept drift are increasingly becoming part of the process (Asliyukse *et al.*, 2025; Kshirsagar *et al.*, 2025). These are some of the insights that we base our work on the literature. We test the techniques in each category which are representative to determine the strengths of each in terms of each other empirically in the same test environment. We explain our methodology in this comparative analysis in the following section.

RESEARCH METHODOLOGY

In order to explore the performance of various techniques of feature representation in detecting email spam, we developed a series of controlled experiments. This section outlines our methodology, such as the datasets that will be used, the feature extraction approaches that will be deployed, the classifiers that will be deployed and the evaluation metrics. We take the form of a conventional supervised learning pipeline in spam classification with close attention to the variation of the feature representation and other factors are held constant at an approximate level of comparison.

Datasets: To make our results comparable to the previous research, we chose two publicly available benchmark datasets commonly used in spam filter research. The first is the Enron-Spam Corpus which is a big set of real email messages of the Enron company and which has been annotated as spam or ham (legitimate) by researchers (Klimt and Yang, 2004). A balanced subsample of the Enron corpus of about 3,000 spam and 3,000 ham messages was used; this was filtered by sampling various mailboxes of different employees to have a variety of topics and writing style. The second data is the CSDMC2010 Spam Corpus (part of the 2010 ECML PKDD Discovery Challenge) which has approximately 4,000 spam and 4,000 legitimate emails and is known to have newer instances of spam such as phishing. These two datasets are complementary to each other: Enron (circa 2001-2002) contains a realistic balance of personal and professional ham, and early-generation spam, whereas CSDMC2010 (circa 2010) contains more modern spam tricks (including hard-to-detect phishing lures). In our context, generally, previous investigations (e.g., Janez-Martino *et al.*, 2023; Kshirsagar *et al.*, 2025) merge several corpora to estimate generalization, although in our case, results will be provided on each dataset individually and on a combined set (encompassing Enron and CSDMC2010), to simulate a larger and more varied email stream (as in Asliyukse *et al.*, 2025). Messages were all processed in a similar way: we removed email header (except subject) so as to concentrate on message content, eliminated HTML tags, and tokenized messages.

We also tested retaining and deleting stop words; in the end, when individual methods (such as BOW) are sensitive to common words, we deleted a set of English common stop words, but in more advanced models (such as BERT embeddings) we retained the text mostly unmodified (since these models are capable of handling common word usage in the context). In the base case we did not lemmatize or stem or anything because we wanted to retain valuable information (some spam filters do stem, however, we did not find it helpful with our dataset and it also complicated the analysis of features).

Feature Representation Techniques: We implemented four distinct feature representation strategies:

1. **Bag-of-Words (BOW):** Each email is expressed as a high dimensional sparse vector of count of words. Each experiment carried a set of 5,000 distinct words (with the removal of stop-words) in the training set that we constructed a vocabulary of. The frequency of each vocabulary word in an email is then stored in the BOW vector of each email. This is the most basic representation and it is a baseline. We also obtained a TF-IDF representation of the counts of the BOWs with each word count being weighted by the inverse document frequency of the word. TF-IDF is said to be a more potent baseline because it shares out the frequent words such as the and dear that do not add much discriminatory information.
2. **Word Embedding Average:** We used 300-dimensional Word2Vec embeddings (placed on the Google News corpus), which are trained to embed words into the continuous vector space. To compute the representation of emails we used the average of the embeddings of all words within the email content (with some basic preprocessing). This gives a dense 300 dimensional representation of the email in the form of a vector. It's intuited that the general topic or semantic information of the email (e.g., an email about a bank account) can be reflected on the average embedding). Although this method sacrifices word order and other smaller contexts, this gives it a more semantic representation compared to BOW. Previous studies have shown that even such simple averaging of word embeddings can be used to improve classification results over raw BOW, due to the introduction of extrinsic semantic knowledge (Bahgat *et al.*, 2018). One more variant we experimented with was TF-IDF weighted average of embeddings, which attaches more weight to rare words during the averaging of word which did not significantly change the results.
3. **Deep Contextual Embedding (BERT):** To obtain a state-of-the-art representation, we used a pre-trained BERT model (BERT-base, uncased) to produce sentence-level email representations. In particular, we input each email (cut to 128 tokens to be efficient) in BERT and pull the embedding of the [CLS] token out of the final layer (which is typically used in classification tasks as a summary one) (Devlin *et al.*, 2019). This yielded a 768-dimensional vector on an email which summarizes contextual data. BERT embeddings, unlike the Word2Vec average, take into account the sequence of words and various meanings of the same word within the context (e.g. distinguishing the meaning of apple as a fruit and another with the word Apple as a company based on the context). In feature extraction, we did not finetune BERT on our spam data, but rather simply used the pre-trained model, which tries to generate embeddings, to test the effect of the representation itself. In the classification step, though, we did one more fine-tuning of another BERT model - see below.
4. **Augmented Feature Set (Hybrid):** We also developed an augmented feature set that took into account textual embeddings as well as a small number of simple structural features based on recent work (Nuha and Lin, 2025; Asliyukse *et al.*, 2025). The features of the structure are: length of email body (number of characters), the number of words, proportion of the uppercase letters (as a percentage of the total alphabetic characters), as well as the number of characters of punctuations. These four features were scaled and combined with the text feature vectors (where such was possible, such as adding the 768-dim BERT feature to

make a 772-dim combined feature, as in the case of Asliyukse *et al.*, 2025). The reason is that a significant number of spam messages contain distinctive stylistic characteristics (such as an abnormally short or long message, over-use of exclamation marks or all-caps to emphasis) which may not be wholly represented by content embeddings. With the inclusion of these indicators, we will increase the rate of detection of lexically advanced but stylistically overt spam..

It should be noted that each representation was computed on the training data and applied to test data in a consistent manner, to avoid any information leakage. For example, the vocabulary for BOW/TF-IDF was built on training emails only, and the IDF values were computed from the training corpus. Similarly, any text normalization (like stop word removal) was done using rules determined a priori, not influenced by test data.

Classification Models: After obtaining feature vectors for emails, we trained machine learning models to classify them as spam or ham. We deliberately use relatively standard classifiers for most representations to ensure that any performance differences can be attributed mainly to the feature representations rather than model complexity. Specifically:

- In the case of the BOW and TF-IDF representations, a Multinomial Naive Bayes (NB) classifier and a Linear SVM were trained. Naive Bayes is an archetype of choices that is efficient and surprisingly effective at spam filtering (Karim *et al.*, 2019). Linear SVM is frequently capable of working with high-dimensional sparse data such as TF-IDF, and can be a powerful baseline (Dada *et al.*, 2019). A logistic regression also provided similar results to SVM and we primarily report SVM due to brevity.
- A Random Forest (RF) classifier and a shallow feed-forward neural network were utilized as a classifier of the averaged Word2Vec embeddings. RF (100 trees) is an effective non-linear classifier that is capable of capturing features interactions; RF has been shown to work well with moderately-sized spam feature sets in the past (AbdulNabi and Yaseen, 2021). Our neural network consisted of one hidden layer with 64 neurons (ReLU activation) and a binary classification layer - effectively to provide an impression whether a simple neural model would be able to make use of the dense vectors compared to RF.
- We tried the BERT contextual embeddings in two modes (a) we use the frozen BERT embeddings as inputs to a classifier (we again tried both RF and a neural network to perform the spam classification task on top of the 768-dim vectors), (b) we fine-tune the BERT end to end on the spam classification task. Fine-tuning implies that we will use the already pre-trained BERT model and then fine-tune it on our labeled spam/ham data (with a classification layer on top) so that it can acquire task-specific insights. The fine-tuning was performed over 2 epochs, and with a learning rate of $2e-5$, which is the suggestion of Devlin *et al.* (2019) in small datasets. The fine-tuned BERT effectively replaces the external classifier and generally performs well when compared to the application of static BERT features with an external classifier, since the model is able to tune the embeddings to attempt to optimise them to the task of spam detection. It is a fact that this method is similar to the recent research such as Nasreen *et al.* (2024) and Asliyukse *et al.* (2025) (only that they occasionally include feature selection or extra layers).
- On the augmented hybrid features (text + structural) considerations, we trained a neural network classifier on the concatenated vector when it is combined with BERT or Word2Vec features (with one hidden layer of 32 units). The reason that we did not rely on NB or SVM with combined features is that these models do not easily accept a combination of dense continuous feature and sparse text feature. The size of the input, in the case of BERT + structural, was 772 dimensions (768 BERT + 4 features) to the neural network..

Lastly, a basic ensemble was also tested: a majority-vote ensemble of three models (NB on TF-IDF, RF on Word2Vec, and fine-tuned

BERT). This was a product of previous research indicating complementary capabilities of various algorithms (Kshirsagar *et al.*, 2025). Our study is not concerned with the ensemble, but we use it to determine whether there is any significant improvement when representations are combined. Evaluation Protocol: We used a standard train-test split protocol. All the data sets were split in a way that 80 percent of the emails were used in training, and 20 percent were reserved in testing. Both splits were made up of equal shares of spam/ham. All hyperparameters of the model (such as the number of trees in RF, regularization in SVM, etc) were optimized using 5-fold cross-validation of the training set of the Enron data and set to remain constant across other experiments to prevent overfitting on the test outcomes. On the one hand, we trained the model using the training data and evaluated using the test set, and we captured the measures such as: accuracy, precision, recall, F1-score, and the false positive rate (false positive - classifying a good email as spam - is especially unwanted in practice). These metrics were calculated individually (i.e. precision is what fraction of mails labeled as spam are actually spam) when required, e.g. precision of spam. In appropriate cases, we also computed the area under the ROC curve (AUC) particularly where the output of a model such as the neural networks or BERT is a probability score - AUC provides an independent of threshold measure of discrimination ability.

Our experiments were repeated 3 times with different random splits (and random initializations of neural models) to ensure the statistical reliability of results and report the average results. The standard deviation was usually minimal (about $\pm 0.5\%$ on accuracy), in which case the results were consistent. We carried out significance tests (two-tailed t-test) to contrast the key models - i.e. between the best traditional model and the best embedding-based model on the same data - using a significance threshold of $p < 0.05$. The experiments were performed on an Intel i7 CPU, 16GB RAM, and NVIDIA Tesla T4 workstation (applied to BERT fine-tuning). The most computationally expensive step was the training of the BERT model (about 10 minutes per epoch on the GPU in our case) as compared to other models which trained within a few seconds. This is the difference that our discussion takes into account with regard to trade-offs of the use of complex representations. With this framework in place, we will be in a direct position to monitor the impact of every feature representation technique on the overall spam classification performance. The results of these experiments will be provided in the following section, where their comparison will be made with the models, and they will be correlated with the existing literature.

RESULTS AND DISCUSSION

In this section, we show our experimental results of the performance of different feature representation methods to use in spam detection. Our first visualization of the results is in Table 1, which presents the summary of the accuracy and critical metrics attained by each combination of methods of representation. Then, we confine ourselves to a comparative discourse, pointing out which techniques were most effective, and explaining them in regard to the nature of spam data. Lastly, we address other learning points like study of errors and effects of concept drift on various models. Comparison of Overall Performance: Table 1 indicates the accuracy of the test-set of various feature representations on the combined dataset (20 percent holdout) which is Enron+CSDMC. In space, we provide Accuracy and F1-score of the spam category (because it is important to detect spam correctly), and the False Positive Rate (FPR) on legitimate emails (because false alarms are very undesirable to allow). All entries are identified with the method of the feature and the classifier. Table 1: The results of different feature representation summaries. The optimal values of each column are underlined. Based on Table 1, we can see that there was a evident ranking of feature representations in terms of accuracy. Naive Bayes on the Bag-of-Words baseline was about 91.8 per cent accurate, which is quite a good result but the lowest of the systems in comparison. TF-IDF followed by a linear SVM was found to achieve an accuracy of 95.6, which validates that re-weighting features based on informativeness and a powerful

classifier can be useful in providing significant improvements (approximately 4 percentage points) and reducing the false positive rate by a factor of half (than plain BOW+NB). It can be matched to the classic works and this finding is the confirmation that our baseline is reasonable (Rapacz *et al.*, 2021; Dada *et al.*, 2019). Using continuous embeddings, the average representation of Word2Vec predicted by a Random Forest with the predictor was 97.4% accurate, surpassing all purely frequency-based representations. This upholds the hypothesis that semantic embeddings can be useful information to differentiate between spam and ham. Specifically, we observed that the Word2Vec-style model was more effective at capturing the presence of vocabulary variation e.g. when an email spam was evading spam-related words but still using similar words or synonyms, the embedding would still tend to put it in the spam category in the vector space. The difference of about $+1.8\%$ improvement over TF-IDF is not a very big though significant difference (t test value played under 0.01, across different runs). Precision and recall were also high with spam F1 of about 97.5, which means that there were very few misses, as well as false alarms. Interestingly, the Random Forest occasionally confused ham emails containing numerous financial terms to be spam, however, we discovered using feature important analysis that the RF was using some of the embedded dimensions which matched finance related words (in spam and in ham) so a few border cases escaped. Those - could be corrected through a more complicated classifier or more context, which is the next observation. The BERT embedding (as a fixed feature vector) also increased the accuracy to 98.5%. The BERT-based neural network the neural network on top of BERT did slightly better than the RF on the same, indicating that the correlation between BERT dimensions and the classes may be somewhat linear/non-linear in a manner that could be better modeled by a small neural net. Fine-tuning BERT by essentially training on a spam classifier resulted in above 99 accuracy. The BERT model was fine-tuned and the accuracy of the final model was 99.1 per cent on the test combined with significantly lowering the false positive error rate to approximately 1.4%. This is an important change: it implies that the model will hardly ever send to the spam legitimate email (only 100 good emails in 100,000 emails, or so, will be wrongly filtered). Practically, this kind of model would have a high use as the complaints made by users usually arise because of the lack of essential (ham) email. The spam F1 of 99.1% shows that it also has a huge proportion of spam that was captured. Such findings are consistent with more recent studies: e.g. Asliyukse *et al.* (2025) state that DistilBERT with added features retrieved the same accuracy on an equally mixed dataset of similarly mixed datasets o.e. 99.4% and our BERT numbers are comparable to that. BERT with feature selection also gave Nasreen *et al.* (2024) more than 99%. The fact that our fine-tuned BERT is able to do this is a strong indication that transformer-based representations are already at the leading edges in this direction.

Also, a small incremental improvement (approximately $+0.2\%$, to 99.3% accuracy) was obtained by adding the four simple structural features to the BERT representation. Although this growth falls within the margin of error, we found that the BERT+Structural model never falsely categorized some of the tricky spam emails as such in the pure BERT model. An example of this would be a single spam email that contained purely an image (no text) passing through the text-based BERT filter (there was not much content to analyze) but our structural characteristics detected an image label and practically no text, which is suspicious - the hybrid model was able to identify the email as spam. This proves the importance of complementary characteristics, which is also stated by Nuha and Lin (2025), who have found out that the combination of content and metadata enhances detection. Though the textual content was the source of virtually all discrimination, such auxiliary cues act as a safety net, especially in the case of outlier content such as spamming with images or extremely brief spam emails. The combination of the various classifiers (NB, RF, BERT) scored approximately 99.0% which was virtually equivalent to the fine-tuned BERT itself. This implies that as soon as one single model (such as BERT) is extremely strong, ensembling does not provide significant returns.

Table 1. Performance of different feature representations for spam detection (combined dataset)

Feature Representation	Classifier	Accuracy (%)	Spam F1-score (%)	False Positive Rate (%)
Bag-of-Words (unigram counts)	Naïve Bayes (Multinomial)	91.8	92.5	7.1
TF-IDF (unigram)	SVM (Linear)	95.6	95.4	4.5
Word2Vec Embedding (avg.)	Random Forest	97.4	97.5	3.2
BERT Embedding (contextual)	Neural Network (1-layer)	98.5	98.6	2.1
BERT (fine-tuned end-to-end)	— (BERT internal)	99.1	99.1	1.4
BERT + Structural features	Neural Network	99.3	99.2	1.2
Ensemble (NB + RF + BERT fine-tuned)	Majority Vote	99.0	99.0	1.5

Table 1: Summary of results for various feature representations. The best values in each column are highlighted in bold

In our example, BERT model was already making the right decision on nearly all of the cases; the cases it was unable to get right were also not always detected by NB or RF, so majority voting was not really a superior solution. At this, ensemble methods were more helpful in the case when we took the Enron dataset separately (which is more diverse in ham). In that case, we have observed the ensemble minimizing false positives slightly as opposed to BERT alone. This suggests that ensembles can be robust in deployments that can change the distribution of data (some users will encounter different distributions of spam/ham). Kshirsagar *et al.* (2025) also discovered that a meta-learner with model combination performed better when datasets were pooled. Error Analysis: We also analyzed a few misclassified examples so as to gain a deeper insight into the weaknesses and strengths of each model. As anticipated, the BOW/NB classifier made errors on most of the spam emails that did not necessarily have spam keywords. An example is the spam with the text "Hi, please review the attached invoice" (a phishing attempt) was rated as ham by NB because the words "please" by itself, as well as the words "review" by themselves, the word "attached" by itself, the word "invoice" by itself are not good spam indicators, and NB assumed independence of features. The BERT model, in contrast, regarded it as spam presumably due to the combination of a generic greeting, a request to open an attachment, and possibly the sender information (the BERT was presented with the entire content and subject in one). This is the importance of context: BOW could not see the forest through the trees in those situations. Conversely, NB was virtually zero false positive on ham that had extraneous keywords of spam in the non-spam settings, but some embedding models misperformed within the initial phases before being fine-tuned. As an example, an Enron ham email talking about a secret investment opportunity in Nigeria (it is a legitimate part of a joke) prompted the Word2Vec+RF model to identify the email as spam because it was semantically similar to typical spam emails.

It was also picked up by the NB model which relies on the presence of known spammy words but once we trained BERT on our data, it learned the style difference (the ham email had an internal thread and corporate signature which are normally absent in spam) and correctly classified it as ham. Such subtle distinction is what the current models are more than capable of doing - they factor in context, sender style, formatting etc. instead of merely keywords. We also saw the concept drift difference between Enron (around 2002) and the spam that are more current. In 2010 spam, the traditional models decreased in accuracy significantly (NB on BOW dropped to about 85% accuracy), but the fine-tuned BERT did not decrease as much (from 99% to 95%). This is consistent with Asliyukse *et al.* (2025) who claimed that performance of the "modern" spam on the context of the static models is deteriorating. The deep and contextual models appear stronger, perhaps due to their generalization of the concept of suspicious content and, perhaps, since the pre-training on BERT provided it with such wide knowledge of language not related to a specific period of spam. However, all models have the advantage of being periodically retrained using recent data to overcome drift. Ideally, a learning system or an ensemble could be put into practice that continuously updates the filter after new trends in spam occur (Karim *et al.*, 2019 discussed adaptive filters in their survey). Statistical Significance: All of the improvements between BOW and TF-IDF to embeddings are statistically significant ($p < 0.05$). Fine-tuned BERT was close to BERT+Structural; in numerous random splits, there was not a constant difference between them. Thus, structural features prove useful but their influence is insignificant in comparison to the strength

of the textual model. The results in terms of the false positive are more favorable - the hybrid model resulted in approximately 20 per cent fewer false positives than BERT by itself in our experiments, which can be critical to user satisfaction. BERT Spam accuracy (the proportion of spam it recognizes as spam) was already 0.995 and hybrid improved it to 0.997 - a slight numerical improvement but possibly a few more legitimate mails were missed. These values are comparable to the best systems in literature (Zouak *et al.*, 2025; Boussaad *et al.*, 2025).

DISCUSSION

Our empirical investigation validates a number of trends investigated by the other researchers. First, emails should be represented with more functionality (whether via embeddings or other features) which enables much more detection. The reason is that the spam of the present era tends to bypass term based filters. As an illustration, numerous phishing emails contain grammatically correct and polite language and company logos - bag-of-words features may fail to help identify the emails, but embedding-based model may detect an insignificant difference or an odd word order, which is associated with a phishing attempt. The excellent results of the BERT model highlight the benefit of context-sensitive representations: the model can successfully comprehend the content of an email on a low-level scale, and as such, it is able to identify even spam of genius. This is at the expense of computational complexity - a learned BERT is much more heavyweight than an NB classifier. In practice, to deploy it, a two-tiered system (Mun *et al.*, 2025) may be used, with a lightweight filter (such as NB or RF on a reduced feature set) identifying blatant spam, and a more expensive model such as BERT analyzing uncertain examples. This would bring about a balance between effectiveness and accuracy.

Secondly, feature selection through metaheuristics (such as GWO, FOA) and ensemble feature methods have also been demonstrated to be useful when the aim is to maximize accuracy without concern about complexity. We did not experiment with FOA and GWO, but the sources (Nasreen *et al.*, 2024; Abualhaj *et al.*, 2025) demonstrate that they can slightly increase performance and make models slimmer by eliminating redundancies. As an example, Nasreen *et al.* (2024) were able to shrink the size of BERT by removing tokens that do not convey information and thereby accelerate the process of inference. In our case, the marginal gain of feature selection was low as BERT is already rather accurate. However, in the case of a Word2Vec or TF-IDF model, these methods may be more fruitful (Ruano-Ordas *et al.*, 2018 discovered that evolutionary algorithms are able to find patterns of words that are indicative of spam). Lastly, our findings support the idea that all metrics do not reveal the entire picture. Although many of the models are accurate, their viability in the real world is based on low false positives. A filter that blocks 1-2 percent of legitimate mails is an issue.

The subtle BERT and our hybrid strategy shone here showing that state-of-the-art representations can reduce false alarms by contextually comprehending representations. Conversely, simple models may inaccurately categorize certain edge-case legitimate emails with spam like words. This subtext is the reason why several email platforms utilize filter layers and enable user reviews (Karim *et al.*, 2019). To compare researchs, however, the F1 and the accuracy itself provide a good feeling of the progress - it is an outstanding

result to move the accuracy range of about 95% to about 99% over the last decade of research, mainly because of the improvement of the representation to better levels. Finally, our findings have been discussed and it can be concluded that the feature representation does play a very vital role in spam detection performance. The trend is evident: the more expressive representations (embeddings, deep context, hybrid features are used), the better spam and ham can be discriminated against. These results... To finalize we can present the findings of our discussion stating that feature representation does really play a decisive role in spam filter success. The successive developments BOW leading to semantic embeddings and then to contextual models were steps in the right direction towards the performance the classifier could have had were it to be perfectly separating spam and legitimate mail. Our empirical findings are in line with what has been observed in the existing literature: modern deep learning-based representations are radically better and more robust in accuracy, yet a wise combination with them (as well as a combination with simpler features) may achieve the highest results.

CONCLUSION

This research paper presents an empirical study of the different feature representation methods of email spam detection starting with the traditional term-frequency models up to the current deep learning embeddings. The findings of the study point out clearly that the effectiveness of spam filters is deeply dependent on the feature representation that is chosen. When combined with algorithms such as Naive Bayes or SVM, traditional representations such as bag-of-words and TF-IDF give a reasonable baseline performance (more than 90 percent accuracy in our experiments). These methods are, however, becoming susceptible to spam content of high sophistication. Including semantic information via word embeddings and context language models, we achieved considerably a better rate of detection. Specifically, contextual embedding models (BERT) could pick up on subtle linguistic features and context to reach 99 percent accuracy in benchmark datasets. This is an enormous improvement on the classical methods and it goes along with a significant decrease in false positives - a critical factor when spam filters are put into practice. Hybrid feature methods (e.g. textual embedding concatenated with email structural features e.g. length of message, use of punctuations) were also investigated. The overall performance of the hybrid models was best in our experiments but with a small margin.

These results indicate that although higher-level text representations are a key contributor to performance improvement, one can find value in multi-faceted sets of features that are able to detect anomalies not related to text content. Also, the feature selection and optimization algorithms (such as GWO or Firefly, which have been demonstrated in related works) can be used to optimize feature sets further, though they might prove especially advantageous in particular cases such as limited computing resources or very unequal data. Some of the practical implications in our analysis are highlighted. To begin with, those companies that want to increase their spam detection methods must invest in current NLP methods - such as the implementation of a fine-tuned transformer model - to remain on top of the changing spam tricks.

Such models are superiorly generalized and thus can deal with new or obscured spam messages better than older filters. Second, a layered filtering approach may be beneficial: basic feature-based simple models can be used to filter the obvious spam with limited resources, and more complex feature-based representation models can be used to examine borderline cases. This two-level mechanism can maximize the efficiency without compromising accuracy (Mun *et al.*, 2025). Third, constant learning is relevant. The optimal feature representation will only deteriorate with time without updating the model to capture emerging trends in spam (concept drift). A spam filter can be kept up to speed as spammers can develop countermeasures to it by feedback loops and periodic retraining on new data.

REFERENCES

- Abualhaj, M., Al-Khatib, S. N., Adel, A. and Anbar, M. (2025). Spam detection boosted by firefly-based feature selection and optimized classifiers. *International Journal of Advances in Soft Computing and its Applications*, 17(1), 45–59.
- AbdulNabi, I. and Yaseen, Q. (2021). Spam email detection using deep learning techniques. *Procedia Computer Science*, 184, 853–858. <https://doi.org/10.1016/j.procs.2021.03.107>
- Alrammahi, A. A. H., Sari, F. A. O., Muhammad, Z. A., Kadhimi, M. N., Al-Shammari, D. and Ibaida, A. (2025). Enhancing spam detection with advanced feature extraction and unsupervised clustering. *International Journal of Information Technology*, 17, 5109–5119. <https://doi.org/10.1007/s41870-025-02723-6>
- Asliyukse, H., Tonkal, O. and Kocaoglu, R. (2025). A comparative evaluation of a multimodal approach for spam email classification using DistilBERT and structural features. *Electronics*, 14(19), 3855. <https://doi.org/10.3390/electronics14193855>
- Bahgat, W., Rady, S., Gad, W. and Moawad, I. (2018). A hybrid approach for email spam filtering using word embedding and classifier ensemble. *IEEE Access*, 6, 28000–28009. <https://doi.org/10.1109/ACCESS.2018.2837038>
- Bacanin, N., Živković, M., Stoean, C., Antonijević, M., Janicijević, S., Šarac, M. and Strumberger, I. (2022). Application of natural language processing and machine learning boosted with swarm intelligence for spam email filtering. *Mathematics*, 10(22), 4173. <https://doi.org/10.3390/math10224173>
- Boussaad, L., Belhadj, A., Thabit, S. and Aoudni, Y. (2025). Spam detection using sentence-BERT with attention-enhanced GRU and SVM. *Applied Computing and Informatics*. (Published online, October 2025). <https://doi.org/10.1108/ACI-02-2025-0035>
- Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O. and Ajibuwa, O. E. (2019). Machine learning for email spam filtering: Review, approaches and open research problems. *Heliyon*, 5(6), e01802. <https://doi.org/10.1016/j.heliyon.2019.e01802>
- Jañez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E. and Alegre, E. (2023). Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach. *Applied Soft Computing*, 139, 110226. <https://doi.org/10.1016/j.asoc.2023.110226>
- Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K. and Alazab, M. (2020). Efficient clustering of emails into spam and ham: A foundational study of a comprehensive unsupervised framework. *IEEE Access*, 8, 154759–154788. <https://doi.org/10.1109/ACCESS.2020.3018488>
- Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K. and Alazab, M. (2019). A comprehensive survey for intelligent spam email detection. *IEEE Access*, 7, 168261–168295. <https://doi.org/10.1109/ACCESS.2019.2954791>
- Kshirsagar, M., Rathi, V. and Ryan, C. (2025). Meta-learner-based frameworks for interpretable email spam detection. *Frontiers in Artificial Intelligence*, 8, 1569804. <https://doi.org/10.3389/frai.2025.1569804>
- Mun, H., Park, J., Kim, Y., Kim, B. and Kim, J. (2025). PhiShield: An AI-based personalized anti-spam solution with third-party integration. *Electronics*, 14(8), 1581. <https://doi.org/10.3390/electronics14081581>
- Nasreen, G., Khan, M. M., Younus, M., Zafar, B. and Saleem, K. (2024). Email spam detection by deep learning models using novel feature selection technique and BERT. *Egyptian Informatics Journal*, 26(6), 100473. <https://doi.org/10.1016/j.eij.2024.100473>
- Nuha, U. and Lin, C.-H. (2025). Investigating the contribution of structural and contextual features for spam detection. *Tehnicki Vjesnik*, 32(3), 958–965. <https://doi.org/10.17559/TV-20240917001995>
- Rapacz, S., Chołda, P. and Natkaniec, M. (2021). A method for fast selection of machine-learning classifiers for spam filtering.

- Electronics*, 10(17), 2083. <https://doi.org/10.3390/electronics10172083>
- Ruano-Ordás, D., Fdez-Riverola, F. and Méndez, J. R. (2018). Using evolutionary computation for discovering spam patterns from email samples. *Information Processing & Management*, 54(2), 303–317. <https://doi.org/10.1016/j.ipm.2018.01.006>
- Tian, Y., Dai, X., Li, Z., Guo, H. and Mao, X. (2025). Improving the accuracy of cybersecurity spam email detection using ensemble techniques: A stacking approach. *PLoS One*, 20(9), e0331574. <https://doi.org/10.1371/journal.pone.0331574>
- Zavrak, S. and Yilmaz, S. (2023). Email spam detection using hierarchical attention hybrid deep learning method. *Expert Systems with Applications*, 233, 120977. <https://doi.org/10.1016/j.eswa.2023.120977>
- Zouak, F., El Beqqali, O. and Riffi, J. (2025). BERT-Graph SAGE: Hybrid approach to spam detection. *Journal of Big Data*, 12, 128. <https://doi.org/10.1186/s40537-025-00746-5>
