Comparative Analysis of Machine Learning and Deep Learning Models for Email Spam Classification Using TF-IDF and Word Embedding Techniques

Kamronbek Yusupov[1], Md Rezanur Islam[1], Ibrokhim Muminov[2], Mahdi Sahlabadi[3], and Kangbin Yim[3](✉)

[1] Department of Software Convergence, Soonchunhyang University, Asan, Korea
{yuskamron,arupreza}@sch.ac.kr
[2] Department of Computer Software Engineering, Soonchunhyang University, Asan, Korea
theibrokhim@sch.ac.kr
[3] Department of Information Security Engineering, Soonchunhyang University, Asan, Korea
sahlabadi@ieee.org, yim@sch.ac.kr

Abstract. This study presents a comparative analysis of machine learning and deep learning using two methods of text data preprocessing for the task of email classification into spam and non-spam. The chosen text preprocessing methods were TF-IDF and Word Embedding. The aim of the experiment was to identify the most effective models for classification tasks and determine the best combination of models with the selected text processing methods. The work used seven deep learning models and seven varieties of machine learning models. The findings revealed that whilst the Word Embedding technique is more appropriate for deep learning models, TF-IDF couples better with machine learning models. Using TF-IDF, Random Forest had the best accuracy among the machine learning models, a score of 0.9787. With scores ranging from 0.9601 to 0.9484, practically all deep learning models showed high accuracy and performance using Word Embedding. Nevertheless, the hybrid CNN-LSTM model efficiently manages classification problems independent of the selected text processing technique.

1 Introduction

In the present digital environment, email is among the most often used and handy channels of communication and information sharing. Its wide use makes it a necessary instrument in both personal and professional spheres. But as email communications have grown, spam [1] that is, unsolicited messages has become far more common. Promotional communications aside, spam seriously threatens people and information systems by including phishing assaults, false schemes, malware distribution, and phishing messages [2].

The issue almost permeates all throughout the globe. With 31.5% of all undesired communications, Russia ranked highest among all the countries in terms of spam emails projected for 2023 [3]. Following with 11.3%, the United States trailed; China came in

L. Barolli (Ed.): BWCCA 2024, LNDECT 231, pp. 114–122, 2025.
https://doi.org/10.1007/978-3-031-76452-3_11

little over 11%. The U.S. sent the most spam emails in a single day nearly eight billion, followed by the Czech Republic and the Netherlands [4] on January 16, 2023. Global email users rose from 3.9 billion to 4.1 billion between 2019 and 2021, and by 2025 they are expected to reach 4.6 billion. Though they have the most internet users, China and India utilize less email than the United States or Germany.

Effective spam categorization now takes front stage as a means of shielding consumers from these dangers. Many techniques and algorithms for spam categorization have been created over the last few decades; each has unique characteristics, benefits, and drawbacks. Long employed to solve this challenge are conventional machine learning techniques such the Naive Bayes (NB) classifier, Support Vector Machine (SVM), and Logistic Regression (LR). These techniques use many algorithms to train models on big datasets and are grounded on statistical analysis. Deep learning techniques have emerged as leading candidates with great efficiency in many natural language processing (NLP) applications as technology and computing capability have developed. Specifically in text classification including spam Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) demonstrate great performance. Particularly pertinent in the face of fast evolving threats, these techniques may capture intricate connections in the data and respond to new kinds of spam. Not less significant are text vectorizing techniques such as Word Embedding and Term Frequency-Inverse Document Frequency (TF-IDF). These techniques enable the numerical vectors created from textual input to be efficiently used to train many models. By use of TF-IDF and Word Embedding, the most important words in context may be found, therefore enabling algorithms to appropriately analyze the text and decide if a message is normal email or spam.

This article aims to provide a comparison of many email spam categorization techniques. The paper will provide a thorough overview of many techniques, including contemporary deep learning techniques and conventional machine learning algorithms. Real-world facts as well as a review of their application in many circumstances will help to determine the efficiency of every approach. Data pretreatment techniques and their fit with the chosen machine learning and deep learning algorithms will be especially under focus. The present work will emphasize the advantages and shortcomings of many approaches as well as point out the most interesting paths for further spam categorization improvement. Researchers as well as practitioners engaged in the creation and use of spam filtering systems in real-world environments might find value in the acquired findings.

2 Related Work

In this section, we will review previous research and work related to the task of spam classification. Over the past decades, numerous methods have been developed for spam classification in email, including both traditional machine learning methods and deep learning techniques. Special attention will be given to algorithms and data preprocessing methods aimed at improving the accuracy of spam email classification.

2.1 Machine Learning Approaches for Spam Email Classification

Mansoor et al. examined machine learning methods for email spam categorization in detail [5]. The paper investigated many techniques including NB, SVM, and Decision Tree (DT). The writers focused especially on the success of these models and investigated their pros and drawbacks. The experimental findings show that high accuracy in spam categorization may be obtained by combining NB and SVM among other methods. However given their complicated design, such combinations are challenging to execute and require large computer resources. For spam email categorization and detection, Siddique et al. suggested a relative study of machine learning and deep learning models [6]. The work examined algorithms including NB, SVM, CNN, and Long Short-Term Memory (LSTM) depending on classification accuracy. To evaluate the models and provide a unique dataset for training, the authors generated LSTM which displayed a high accuracy of 98.4% among other models. Training LSTM, they observed, calls for much more computer resources and time than models of CNN, NB, and SVM. Using Word Embedding, Somesha et al. presented a comparative study of machine-learning techniques for phishing email categorization [7]. Applied data preprocessing techniques like TF-IDF, Count Vectorization, Word2Vec, and FastText, the research examined machine learning models including Random Forest (RF), DT, SVM, XGBoost, and LR. FastText and RF taken together produced a high accuracy of 99.50% in phishing email categorization, according to the findings. With accuracy between 95% and 99%, other models also proved to be successful. Still, reaching such high performance calls for rigorous model parameter adjustment over several datasets. For purposes of spam email categorization and detection, Kang et al. performed a comparative study of machine learning models [8]. Using the n-gram approach, the paper evaluated and contrasted five forms of machine learning models: NB, DT, RF, SVM, and K-Nearest Neighbors (KNN). Training and model testing came from two well-known Kaggle datasets. On both sets, the SVM model proved to have the best accuracy 97.5%. Authors pointed out, meanwhile, that the features of certain datasets determine how well each model performs.

2.2 Deep Learning Approaches for Spam Email Classification

Ghourabi et al. suggested a mixed deep learning model [9] to distinguish regular from spam communications. The paradigm aggregates LSTM and CNN techniques. Using an improved form of TF-IDF known as Term Frequency-Average Document Frequency (TF-ADF), which takes term frequency into account for a more accurate portrayal of the word value, their research. According to tests, the suggested model achieved 98% in spam categorization. The approach does, however, need a lot of processing capacity and TF-ADF tuning is challenging and requires careful parameter selection for optimum operation. Bansal et al. proposed a hybrid approach founded on artificial neural networks (ANN) [10] for spam message classification. The article tested several machine learning methods like NB and XGBoost against the model. TF-IDF transformed text input into numerical vectors, therefore simplifying model training. The suggested model exceeded NB and XGBoost in important metrics with an accuracy of 97.5%; NB and XGBoost had accuracies of 87.8% and 92.7% respectively. One of the negatives of ANN, meanwhile, is the need for a lot of data for good training. In the lack of enough data, the model could suffer poor performance or overfit. Transfer Learning approaches might be seen as [11] a solution to this dilemma. For spam email categorization, Debnath et al. put up a deep learning-based model [12]. Their technique integrates LSTM, Bidirectional LSTM (BiLSTM), and Bidirectional Encoder Representations from Transformers (BERT) unlike earlier research. In their investigation, the authors used Word Embedding and TF-IDF to enhance data quality using text preparation. Experimental results showing out of LSTM and BiLSTM demonstrated that the BERT model achieved the greatest accuracy, 99.14%. For comparable tasks, they also assessed machine learning models with models showing accuracy ranging from 93% to 97%). Of the models tested SVM, KNN, DT, LR, RF, and Multinomial Naive Bayes (MNB) had the greatest accuracy. Furthermore, lacking in MNB are various assumptions of feature independence, which might compromise accuracy. At the same time, BERT requires more training time and computer resources than other models.

3 Methodology

In this section, we will provide a brief overview of the models we used, their architecture, and the differences between them. We will discuss the strengths and weaknesses of each model. At the end of this section, we will also review the data preprocessing methods, such as TF-IDF and Word Embedding, that we employed to enhance the training of our models.

3.1 Overview of Selected Machine Learning and Deep Learning Models for Spam Email Classification

This work developed an email spam classifier using two techniques: deep learning and machine learning. The first method included choosing machine learning models like LR, SVM, RF, Gradient Boosting (GB), DT, NB classifier, and KNN. The second method used deep learning models like Multilayer Perceptron (MLP), LSTM, RNN, Gated Recurrent Unit (GRU), CNN-LSTM, Attention model, and Transformer.

LR is a linear model used to predict if an email falls into the spam or non-spam categories. Considered successful for tasks with high-dimensional characteristics, the SVM model discovers a hyperplane separating the two classes of data. Training each tree on a random selection of data, the RF model reduces overfitting risk on complicated data. With each next phase, GB corrects the mistakes of past models, hence enhancing accuracy. DT determines the final class by building a chain of branching depending on characteristics, therefore attaining great accuracy. Based on Bayes' theorem and the supposing of feature independence, the NB classifier is a basic method. This method individually assesses every word in the text for likelihood. KNN uses the distance to the closest neighbors to establish the class. This approach does, however, have flaws including susceptibility to noise and great computational expense.

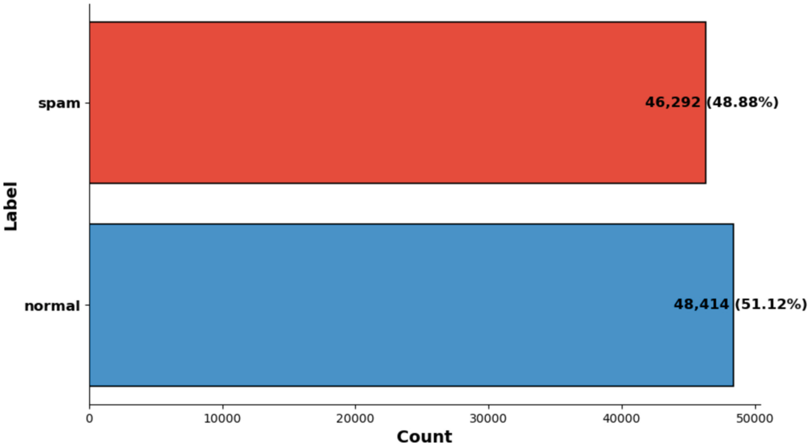

Fig. 1. Distribution of Normal and Spam Emails in the Combined Dataset

One of the first forms of neural networks, multilayer perceptron (MLP) has numerous hidden layers where every neuron implements a nonlinear activation mechanism. Simple RNN, LSTM, and GRU are three primary RNN variants. SimpleRNN was the first version applied as it considers text sequence dependencies and order. However, this model finds the vanishing gradient issue in its design, which makes learning long-term relationships challenging. LSTM and GRU were created to solve this problem. These two designs run on similar ideas and have the same aims. LSTM and GRU vary mostly in their architecture: Three gates in LSTM regulate data flows, guiding information loss and passing on decisions. Although it uses a lot of computer resources, this kind of method excels on difficult problems. Having only two gates reset and update GRU is a simplified kind of LSTM. Faster, simpler, and more efficient than LSTM this kind of model is. CNN-LSTM is a hybrid model wherein CNN generates features from the data and then forwards them to LSTM for further processing. Modern methods of text processing chores include Attention models and Transformers. By letting the model concentrate on crucial elements of the input data, the attention mechanism helps to clarify the context. Processing data sequences, the Transformer model concurrently focuses on the element interactions.

3.2 Text Pre-processing Techniques: TF-IDF and Word Embedding Implementation

The research made use of a dataset from Kaggle [13], which already included the required labels. Two identical datasets were discovered [14], subsequently merged, and ready for model training. With a total of 94,706 emails, the final merged dataset consisted of two columns: text and label. Of these, 48,414 were labeled as normal as indicated in Fig. 1 and 46,292 were labelled as spam. Data preparation finished; thereafter, data preprocessing started. Model training depends much on data pretreatment as it helps avoid overfitting or data interpretation errors. This paper studied two primary text data preparation methods, TF-IDF and Word Embedding, to produce a robust spam message classifier. This comparison aimed to find which method suited certain models of deep learning and machine learning. These methods taken together provide a whole approach to explain text data, therefore helping the models to more accurately classify emails as spam or non-spam.

Table 1. Comparative Accuracy of Machine Learning and Deep Learning Models Using TF-IDF and Word Embedding for Spam Email Classification

Type	Models	TF-IDF	Word Embedding
Machine Learning	LR	0.9642	0.8758
	SVM	0.9684	0.8786
	RF	0.9787	0.9491
	GBM	0.9198	0.8997
	DT	0.9401	0.8671
	NB	0.9313	0.8225
	KNN	0.8370	0.9373
Deep Learning	MLP	0.9738	0.9601
	SimpleRNN	0.8658	0.8871
	LSTM	0.8908	0.9725
	GRU	0.8906	0.9715
	CNN-LSTM	0.9764	0.9794
	Attention Model	0.8855	0.9692
	Transformer Model	0.8752	0.9608

TF-IDF [15] is the first technique we used based on the idea of evaluating the relevance of every phrase in the text concerning the whole corpus of documents. Two crucial phases are involved in this approach: computing term frequency (TF) and assessing inverse document frequency (IDF), thereby displaying the general word rarity. This method lets the model focus on words that could be unique and have critical relevance or importance for a certain text, therefore excluding common keywords that lack relevant information for categorization. This approach helps the model concentrate on certain keywords most often present in the dataset during training. For our dataset, such words were “free” and “win,” respectively. This method is thought to be effective when the dataset comprises several essentially identical or frequent words.

The second method is word embedding [16] vectors words. From text data reflecting the semantic relationships among words in a particular manuscript, this method produces dense numerical vectors. This method helps one consider the context of word usage unlike TF-IDF, which views every word individually. Given its potential, word embedding is seen as a great tool for text analysis. Learning on large text corpora, the method presents words as vectors in a multidimensional space. These vectors help machine and deep learning models not only to concentrate on particular words but also to grasp their connections, therefore enhancing the general comprehension of the text. With the Word Embedding approach, for instance, the model can identify the resemblance between the phrase “offer” and “promotion”, or “sales” even if they do not necessarily fit in the same context but can be utilized in spam messages.

4 Experimental Evaluation

In this section, we will conduct a comparative analysis of machine learning and deep learning models for the task of classifying emails into spam and regular emails. The main goal of the present experiment was to find the most efficient model in conjunction with many approaches of data preparation. Two techniques to text data representation were used in order to reach this aim: TF-IDF and Word Embedding. TF-IDF was the first kind of preprocessing used to translate books into numerical vectors reflecting word significance. Word Embedding was used to generate dense vector representations of words, unlike TF-IDF, therefore enabling the model to incorporate semantic links between words. Using the methods, several machine learning models, such as LR, SVM, RF, GBM, DT, NB, KNN, as well as deep learning models, such as MLP, SimpleRNN, LSTM, GRU, CNN-LSTM, Attention, and Transformer Model, were trained. The experimental results varied depending on the chosen data preprocessing method. For better clarity, the results are presented in Table 1 and Fig. 2.

The results of the first experiment, where TF-IDF was used, showed that the RF model demonstrated high accuracy in spam email classification tasks, achieving an accuracy of 0.9787 as shown in Fig. 2. In the case of deep learning, the best result was achieved with the CNN-LSTM model, with a classification accuracy of 0.9764. However, almost all models using TF-IDF showed promising results, except for the KNN machine learning model, which proved incompatible with TF-IDF, showing accuracy ranging from 0.7842 to 0.8370.

In the second experiment, where the Word Embedding method was used, the high results achieved with TF-IDF on machine learning models showed a decrease in accuracy by 5–8%. However, KNN, which showed low accuracy when using TF-IDF, demonstrated high accuracy when using the Word Embedding method, reaching 0.9373. Among other models, the most effective was the CNN-LSTM hybrid, which achieved spam classification accuracy up to 0.9794 as shown in Fig. 2. It is also worth noting that the use of Word Embedding in deep learning models showed higher results compared to TF-IDF. This is especially true for LSTM, GRU, Attention, and Transformer Models, which showed accuracy ranging from 0.9608 to 0.9725, whereas with TF-IDF, the accuracy ranged from 0.8752 to 0.8908.

The experimental results demonstrate that the choice of the right model and preprocessing method plays a key role in building an effective model for any task and has a significant impact on the quality of classification. Based on the results, machine learning models showed good compatibility with the TF-IDF method compared to Word Embedding. However, in the case of deep learning models, on the contrary, Word Embedding proved to be a more compatible method for spam classification tasks compared to TF-IDF. Despite this, a model compatible with both data preprocessing methods was found the CNN-LSTM, which showed high efficiency in both cases, indicating that such hybrid approaches can effectively handle various tasks.

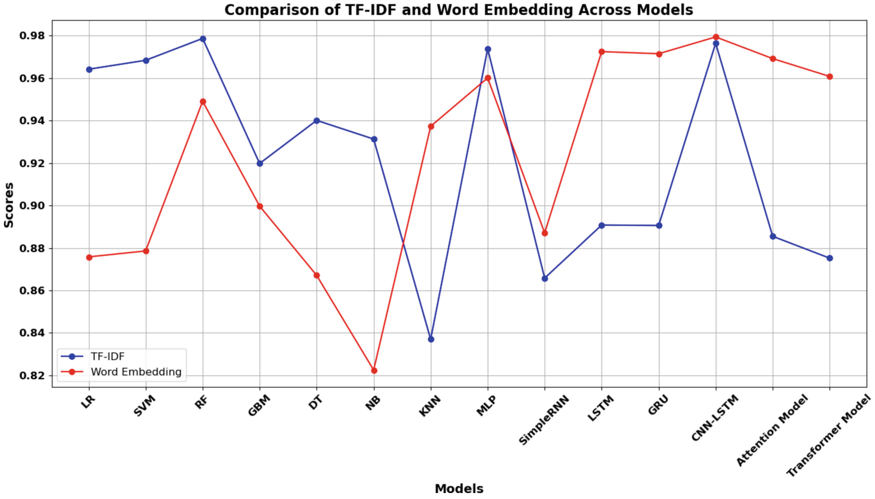

Fig. 2. Comparison of TF-IDF and Word Embedding Across Models

5 Conclusion

The goal of this work was to categorize emails into spam and legitimate emails by means of a comparison between machine learning and deep learning models. Two text data preparation techniques, TF-IDF and Word Embedding, were also compared in this work. We aimed to find which techniques complement one another and which fit certain models more suitably. Study results suggest that Random Forest shows high accuracy among machine learning models using TF-IDF. Deep learning models incorporating Word Embedding such as LSTM, GRU, Attention, and Transformer achieved remarkable success in spam email classification difficulties. Interestingly, hybrid models such as CNN-LSTM also showed remarkable accuracy in utilizing both data preparation methods regardless of the chosen method. Therefore, in practical situations, hybrid methods such as CNN-LSTM may efficiently manage spam email categorization chores. This is confirmed by our experiment, in which CNN-LSTM demonstrated high efficiency. Based on the results of our experiments, it should also be noted that TF-IDF methods are better suited for machine learning models, while Word Embedding works better with deep learning models. In the future, our research will focus on identifying more hybrid models and determining the best combinations of text data preprocessing methods for these models.

Acknowledgments. This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Convergence security core talent training business support program (IITP-2024-2710008611) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

References

1. Keskin, S., Sevli, O.: Machine learning based classification for spam detection. Sakarya Univ. J. Sci. **28**(2), 270–282 (2024)
2. Abdillah, R., et al.: Performance evaluation of phishing classification techniques on various data sources and schemes. IEEE Access **11**, 38721–38738 (2022)
3. Kaspersky Lab.: Leading Countries of Origin for Unsolicited Spam E-mails in 2023, by Share of Worldwide Spam Volume. Statista, Statista Inc., 7 Mar 2024. https://www.statista.com/statistics/263086/countries-of-origin-of-spam/
4. Cisco Talos Intelligence Group. Daily Number of Spam Emails Sent Worldwide as of January 2023, by Country (in Billions). Statista, Statista Inc., 16 Jan 2023. https://www.statista.com/statistics/1270488/spam-emails-sent-daily-by-country/
5. Mansoor, R.A.Z.A., Jayasinghe, N.D., Muslam, M.M.A.: A comprehensive review on email spam classification using machine learning algorithms. In: 2021 International Conference on Information Networking (ICOIN). IEEE (2021)
6. Siddique, Z.B., et al.: Machine learning-based detection of spam emails. Scientif. Programm. **2021**(1), 6508784 (2021)
7. Somesha, M., Alwyn, R.P.: Classification of phishing email using word embedding and machine learning techniques. J. Cyber Secur. Mobility **11**(3), 279–320 (2022)
8. Kang, G., et al.: The comparison of machine learning methods for email spam detection. In: International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. Cham: Springer Nature Switzerland (2023)
9. Ghourabi, A., Mahmood, M.A., Alzubi, Q.M.: A hybrid CNN-LSTM model for SMS spam detection in Arabic and English messages. Future Internet **12**(9), 156 (2020)
10. Bansal, C., Sidhu, B.: Machine learning based hybrid approach for email spam detection. In: 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). IEEE (2021)
11. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. J. Big Data **3**, 1–40 (2016)
12. Debnath, K., Kar, N.: Email spam detection using deep learning approach. In: 2022 international conference on machine learning, big data, cloud and parallel computing (COM-IT-CON), vol. 1. IEEE (2022)
13. Venky73. Spam Mails Dataset (2022). Kaggle, https://www.kaggle.com/datasets/venky73/spam-mails-dataset
14. UCI Machine Learning Repository. *SMS Spam Collection Dataset* (2019). Kaggle, https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset
15. Minhaz Hossain, S.M., Aashiq Kamal, K.M., Sen, A., Sarker, I.H.: Tf-idf feature-based spam filtering of mobile sms using a machine learning approach. In: Siddique, N., Arefin, M.S., Shamim Kaiser, M., Kayes, A.S.M. (eds.) Applied Intelligence for Industry 4.0, pp. 162–175. Chapman and Hall/CRC, New York (2023). https://doi.org/10.1201/9781003256083-13
16. Srinivasan, S., et al.: Spam emails detection based on distributed word embedding with deep learning. In: Maleh, Y., Shojafar, M., Alazab, M., Baddi, Y. (eds.) Machine intelligence and big data analytics for cybersecurity applications, pp. 161–189. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-57024-8_7