ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 12, Issue 5, May 2025

Multi-Class Email Categorization in Enterprise Environments: A Study of Traditional SVM and Transformer-Based XLNET Models

[1] Vishal Tyagi, [2] Sarthak Marwah, [3] Dr. S. Nagadevi, [4] Dr. Sindhuja M, [5] Anand M

[1] [2] [3] [4] [5] Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Tamil Nadu, India

Corresponding Author Email: [1] Vs7747@srmist.edu.in, [2] Sr4536@srmist.edu.in, [3] nagadevs@srmist.edu.in, [4] sindhujm1@srmist.edu.in, [5] anandm8@srmist.edu.in

Abstract— The increasing volume of electronic communication makes effective email management essential for businesses. This research utilizes both conventional and sophisticated machine learning techniques to analyze multi-class email classification. The baseline is a Support Vector Machine (SVM) model with an F1-score of 62.01% and an accuracy of 67.07%. The Mean Squared Error (MSE) is 2.8710, the Root Mean Squared Error (RMSE) is 1.6944, the Mean Absolute Error (MAE) is 0.9885, the Standard Deviation (SD) is 1.6741, the Correlation Coefficient (R) is 0.4144, and the Coefficient of Determination (R^2) is 0.4320, which are significant performance metrics for the SVM model. The model is suitable for structured email categorization tasks owing to its adequate performance and reasonable consistency. Conversely, an XLNet-based Large Language Model (LLM) methodology is optimized on the identical dataset to leverage contextual embeddings for enhanced classification. The LLM demonstrates an accuracy of 69.98% (rounded to 70%) and an F1-score of 54.89%, surpassing the SVM in F1 performance while exhibiting slightly superior accuracy. With statistical values being at 2.0538 for MSE, 1.4331 for RMSE, 0.7142 for MAE, 1.4287 for SD, 0.4197 for R, and 0.0244 for R, the LLM model is statistically significant. Even though LLM has more contextual awareness compared to the SVM model, it seems to perform similarly. Comparing to Support Vector Machines (SVMs) and other low-complexity ML approaches for structured categorization, LLMs generally use relatively complex contextual embeddings. As the adoption of e-communication continues to rise, the need of effective email management, especially for businesses is growing. The comparative analysis highlights the strengths and limitations of both approaches, offering insights into their scope, applications and deployment scenarios.

Keywords— Email Classification, Machine Learning, Support Vector Machine, Large Language Model, Natural Language Processing.

I. INTRODUCTION

Email is still an important way to communicate in this digital age, but the number of messages is growing, which makes it harder to organize and sort them effectively. Keyword-driven and rule-based traditional methods are simple, but they can't handle the complex and changing nature of modern communication and can't be used on a large scale. They often result in disorganized inboxes, incomplete or missed information, and in turn also reduced productivity since they are unable to identify the semantic variances necessary for precise email labeling. Using machine learning and artificial intelligence has significantly changed how emails are automatically categorized.

Standard machine learning techniques, especially Support Vector Machines (SVMs), are widely used because they consistently work well and are fast to compute in structured classification tasks. But the fact that they still don't fully understand the nuances of unstructured writing in context is a big problem. Large Language Models (LLMs) like XLNet have transformed text separation, nevertheless, by using their awareness of deep semantic processing and contextual significance. These very sophisticated models are computationally demanding and not totally dependable even if they provide efficient and desirable solutions for managing large email volume. This work uses a thorough strategy to close the gap between computational efficiency and contextual accuracy by assessing and contrasting SVM and LLM based email categorization systems. First, an SVM model is used as a baseline, including necessary preprocessing operations including tokenization, text cleaning, and TF-IDF feature extraction to translate textual input into numerical representations.

Supported by performance data comprising MSE of 2.8710, RMSE of 1.6944, MAE of 0.9886, SD of 1.6741, R of 0.4144, and R^2 of 0.4320, the SVM model shows modest efficacy with an accuracy of 67.07% and an F1-score of 62.01%. Then, to use enhanced contextual embeddings, an XLNet-based LLM is fine-tuned on the same data. With performance measures including MSE of 2.0538, RMSE of 1.4331, MAE of 0.7142, SD of 1.4287, R of 0.4197, and R^2 of 0.0244 the LLM achieves an accuracy of 69.98% approximated by 70% and an F1-score of 54.89%. This comparison shows the dependability and light weight of SVM as well as the possibilities of LLMs for improved contextual understanding. By means of enhanced categorization accuracy and contextual awareness, the results

ISSN (Online) 2394-2320
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Volume 12, Issue 5, May 2025

help to build intelligent email management systems, thereby streamlining organizational processes.

The organization of this work follows: Section II addresses related tasks; Section III goes into great length on the suggested approach. Experimental data and performance measures are given in Section IV. Section V ends the analysis with important results and possible future developments.

II. RELATED WORKS

The study by MCNN-LSTM: Combining CNN and LSTM to Classify Multi-Class Text in Imbalanced News Data (2021) introduces a hybrid model that leverages CNN and LSTM networks for multi-class text classification, particularly focusing on imbalanced datasets. The model demonstrates improved accuracy by capturing spatial and sequential dependencies within text data.

Nowak, E., Vidal, E., & Zoghby, C. [1] (2021) in their paper Email Classification Using LSTM: A Deep Learning Technique explore the effectiveness of Long Short-Term Memory (LSTM) networks in classifying emails. The study highlights LSTM's ability to capture long-term dependencies in sequential data, making it suitable for text classification tasks.

Saxena, K., & Bhattacharyya, A. [2] (2017) in E-mail Classification with Machine Learning and Word Embeddings for Improved Customer Support investigate the use of word embeddings in combination with machine learning models to enhance email classification. Their research focuses on improving classification accuracy in customer support applications.

Banday, M. T., & Sheikh, S. A. [3] (2014) present Realization of Microsoft Outlook® Add-In for Language-Based E-Mail Folder Classification, which discusses the development of an Outlook add-in for classifying emails based on language. The study serves as a foundation for topic-based and priority-based email classification methodologies.

Sergio Rojas-Galeano [4] (2024) in Zero-Shot Spam Email Classification Using Pre-trained Large Language Models investigates the application of zero-shot learning for spam email classification. The study suggests that more research is required to adapt zero-shot methods for multi-class email classification, particularly in evaluating large models like GPT-4 for general email categorization tasks.

Nowak, E., Vidal, E., & Zoghby, C. [5] (2021) in Email Classification Using LSTM: A Deep Learning Technique identify the limitations of LSTM in multi-class email classification, particularly the lack of domain-specific feature considerations that could enhance classification performance.

Maxime Labonne & Sean Moran [6] (2023) introduce Spam-T5: Benchmarking Large Language Models for Few-Shot Email Spam Detection, which benchmarks the effectiveness of large language models (LLMs) for email spam detection. However, the study primarily focuses on binary classification (spam vs. non-spam), highlighting the need for research in non-binary email classification tasks.

The paper A Comparative Analysis of SVM, LSTM, and CNN-RNN Models for the BBC News Classification [7] (2021) compares multiple machine learning and deep learning models for news classification. The study suggests that hybrid models combining SVM and LSTM could provide more robust multi-class classification across diverse domains, including email classification.

III. IMPLEMENTATION

This section outlines the step-by-step implementation of the email classification system, focusing on the architectural design and processes underlying the Support Vector Machine (SVM) model and the fine-tuning of Large Language Models (LLMs).

A. SVM Approach

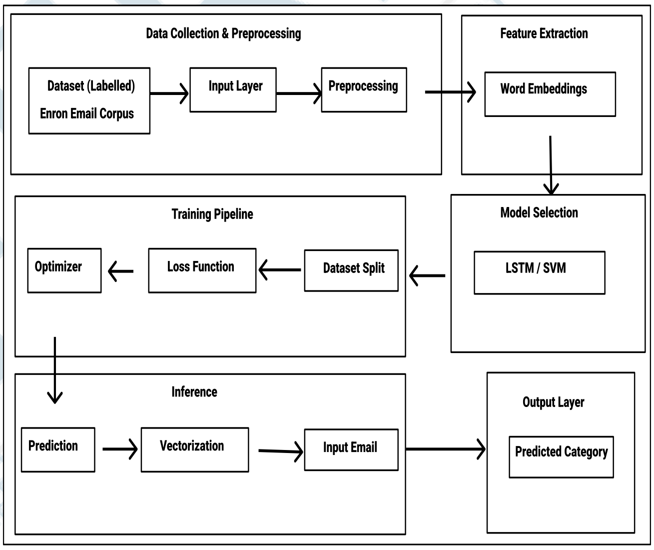

Fig 1: SVM Architecture Diagram

The first architecture is based on an SVM approach, made specifically for email classification. The process starts with data collection and preprocessing, where a raw email data is gathered from a labeled dataset, such as the Enron Email Corpus. The input layer does the processing of this data by applying standard cleaning techniques, including the removal of punctuations, stopwords, and other unnecessary tokens, followed by tokenisation to prepare the text for feature extraction. Feature extraction is subsequently performed using word embedding techniques such as Word2Vec or Term Frequency-Inverse Document Frequency (TF-IDF), which converts data in the form of text into numerical representations which are suitable for machine learning algorithms.

The classification model is built using Support Vector Machine (SVM). SVM acts as a baseline classifier, mapping all the input features to their corresponding email categories through a hyperplane-based decision boundary. The training

ISSN (Online) 2394-2320
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Volume 12, Issue 5, May 2025

pipeline involves splitting the whole dataset into training and testing sub-datasets, the model is trained using kernel-based techniques to maximize class separability and distinguish in feature space.

Once trained, SVM model produce classified email outputs. The output layer assigns emails to predefined categories, demonstrating the effectiveness of both approaches in automating email classification tasks

B. LLM-Based Approach

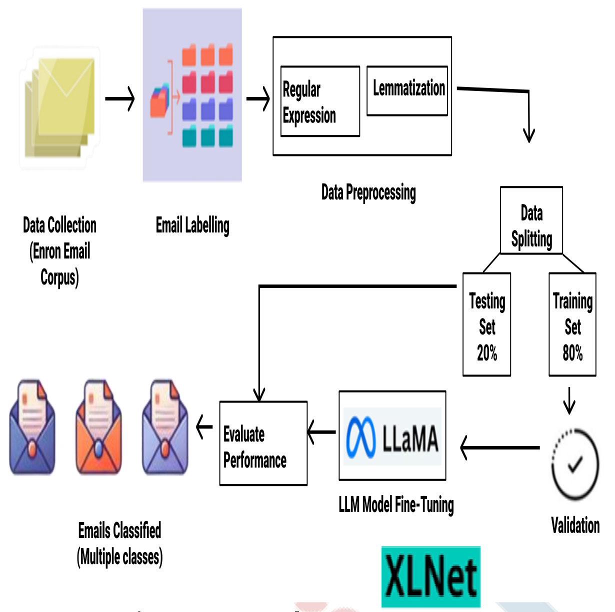

Fig 2: LLM Architecture Diagram

The implementation starts with data collection, where raw email datasets are collected as an input for the classification pipeline. These emails are then systematically labeled into predefined categories as per user requirements such as HR, Finance, and Business, forming a dataset that is structured for supervised learning. The labeled emails undergo massive preprocessing, including the removal of unwanted characters and special symbols through regular expressions, as well as lemmatization to normalize words to their root forms. This step ensures uniformity in textual data and enhances model interpretability.

Once the preprocessing of data is done, the dataset is then partitioned into training (80%) and testing (20%) sets to facilitate model training and evaluation. The classification task is performed by fine-tuning the pre-trained models such as Bidirectional Encoder Representations from Large Language Model XLNet on the training dataset. Fine-tuning of these models basically ensures that their embeddings align with the specific characteristics of the email dataset, allowing them to generalize better against the real-world classification tasks.

During training, validation is conducted to assess the generalization ability of the models and prevent overfitting. Once the training is complete, the model is then evaluated using our pre-determined standard performance metrics, which include accuracy, precision, recall, and F1-score, to measure classification effectiveness. Finally, our email classification system generates categorized outputs, demonstrating high accuracy and scalability in processing and categorizing emails based on content as well as context.

IV. RESULTS AND DISCUSSION

Both preprocessing and organization of the email classification dataset helped to improve the performance of the model. Indeed, the email examples were categorized into predefined groups that allowed both Large Language Model (LLM) and Support Vector Machine (SVM) based approaches to learn rather effectively. The dataset was arranged in a structured way including input text, extracted features, and matched labels thereby ensuring compatibility with both machine learning and deep learning models.

The raw email dataset for the SVM-based approach was tokenized following cleansing against punctuation, stop-words, and other superfluous characters. By means of numerical representations, Term Frequency-Inverse Document Frequency (TF-IDF) was also used to derive characteristics. The SVM model was trained with kernel-based methods in order to translate our input features into the suitable categories. With an F1-score of 62% and an accuracy of roughly 67%, the model displayed modest efficacy in the categorization of structured emails.

In the XLNet approach, the dataset underwent advanced text normalization techniques, also including lemmatization and special character removal, to prepare the text for model training. Fine-tuning of the model was performed using XLNet model, which was specifically adapted for the task of email classification. By means of an adjustable learning rate, the model was tuned to match all pre-trained contextual embeddings with our segregation aim. To guarantee strong performance evaluation, the training procedure has split its 80-20 dataset both for testing and training accordingly. With an accuracy of about 70% and an F1-score of 55%, the model somewhat improved over the already used SVM technique.

Standard metrics—including accuracy, precision, recall, and F1-score—evaluated the categorization performance. Whereas the classification performance measures are compiled in Fig. 3, the accuracy and loss curves for both models are shown in Fig. 4 and Fig. 5. The results show that although providing enhanced contextual awareness and adaptability, the XLNetbased method shows equivalent performance to the SVM-based method. Comparatively analyzing the performance of both models in Fig. 6 highlights their various strengths and constraints in multi-label email categorizing.

To further facilitate comparison, we present the performance metrics in a tabular format, accompanied by a bar graph and a separate line graph depicting the XLNet model's performance over epochs.

ISSN (Online) 2394-2320
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Volume 12, Issue 5, May 2025

Model	Accuracy	F1 Score
SVM	67%	62%
XLNet	70%	55%

Fig 3: Comparison Table of Accuracy and F1 Score between SVM and XLNet Models

The bar graph and line graph are presented in Fig. 4 and Fig. 5, respectively, for better visualization.

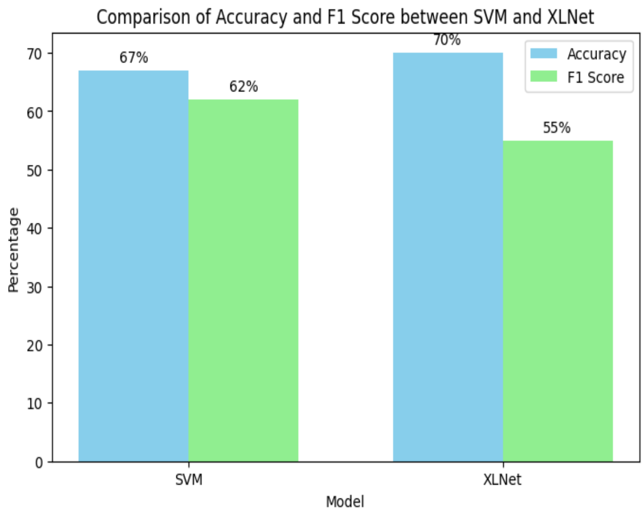

Fig 4: Comparison of Accuracy and F1-Score between SVM and XLNet Models

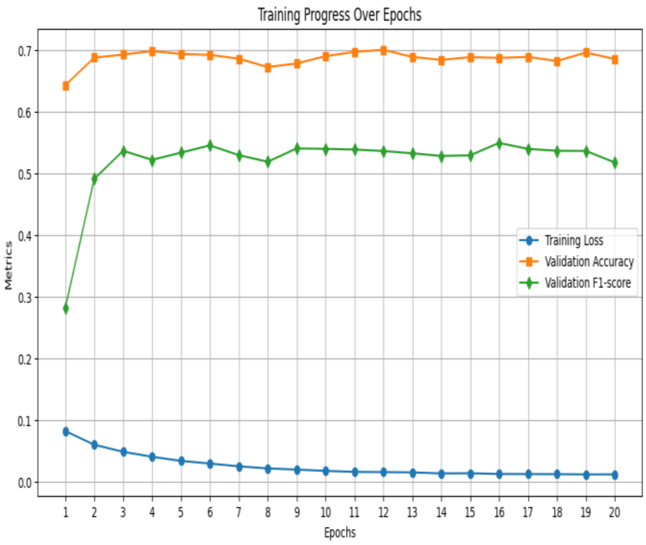

Fig 5: Training Progress of XLNet Model over Epochs

To provide a comprehensive comparison between the SVM and XLNet approaches, key performance metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Standard Deviation (SD), Correlation Coefficient (R), and Coefficient of Determination (R^2) were evaluated. The comparison table presented in Table I clearly outlines the differences in these metrics between the two models, highlighting the relative strengths and weaknesses of each approach.

Metric	SVM	XLNet
MSE	2.8710	2.0538
RMSE	1.6944	1.4331
MAE	0.9885	0.7142
SD	1.6741	1.4287
R	0.4144	0.4197
R^2	0.4320	0.0244

Fig 6: Performance Metrics Comparison between SVM and XLNet Models

The visual representation of these metrics is shown in Fig. 7, where the bar graph illustrates the performance differences between the SVM and XLNet models. The analysis reveals that XLNet exhibits lower error rates (MSE and RMSE) and higher consistency (lower SD) compared to SVM, while both models show comparable correlation coefficients.

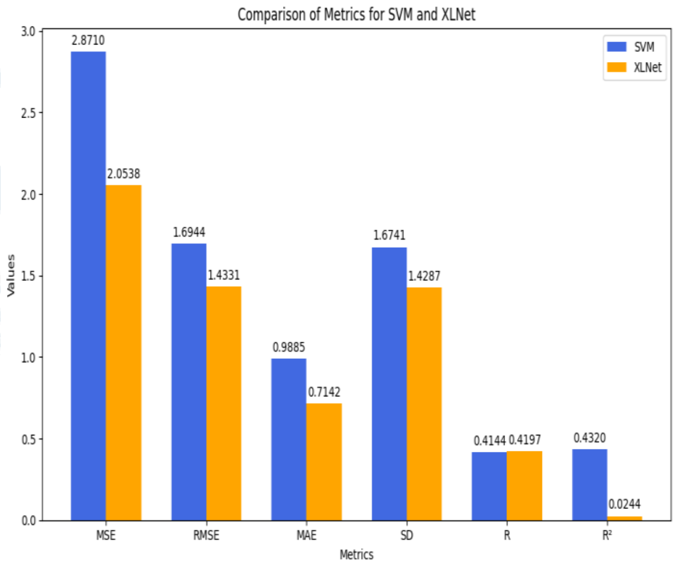

Fig 7: Comparison of Metrics for SVM and XLNet

V. CONCLUSION

This study demonstrates the potential of advanced machine learning and deep learning techniques in optimizing email management through multi-class classification. By implementing Support Vector Machines (SVM) and fine-tuning Large Language Models (LLMs), the proposed approach effectively addresses the challenges associated with categorizing emails into predefined classes. The architecture integrates robust data preprocessing methods with state-of-the-art models, ensuring high classification accuracy and improved email organization.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 12, Issue 5, May 2025

The comparative analysis between traditional machine learning models such as SVM and modern LLMs highlights the significant advancements in natural language processing. The results indicate that LLMs, particularly XLNet exhibit superior flexibility, scalability, and contextual understanding, making them more effective for multi-class classification in diverse organizational environments. This research further underscores the critical role of data preprocessing, model selection, and performance evaluation in enhancing classification outcomes, contributing to improved email management solutions for both business and research applications.

REFERENCES

[1] Saxena, K., & Bhattacharyya, A. (2017). "E-mail Classification with Machine Learning and Word Embeddings for Improved Customer Support." Neural Computing and Applications.

[2] Banday, M. T., & Sheikh, S. A. (2014). "Realization of Microsoft Outlook® Add-In for Language-Based E-Mail Folder Classification."

[3] Rojas-Galeano, S. (2024). "Zero-Shot Spam Email Classification Using Pre-trained Large Language Models." arXiv, Cornell University.

[4] Nowak, E., Vidal, E., & Zoghby, C. (2021). "Email Classification Using LSTM: A Deep Learning Technique." IEEE Xplore.

[5] Labonne, M., & Moran, S. (2023). "Spam-T5: Benchmarking Large Language Models for Few-Shot Email Spam Detection." arXiv, Cornell University.

[6] A Comparative Analysis of SVM, LSTM, and CNN-RNN Models for the BBC News Classification. (2021). ResearchGate.

[7] MCNN-LSTM: Combining CNN and LSTM to Classify Multi-Class Text in Imbalanced News Data. (2021). ResearchGate.