Journal of Information Hiding and Multimedia Signal Processing ©2025 ISSN 2073-4212
Ubiquitous International Volume **16**, Number **2**, June **2025**

AEEM: A Deep Learning Approach to Automated Educational Email Management

Thu Duong[1,2], Anh Tran[1,2], Vy Dang[1,2], Tu Tran[1,2], Hoanh-Su Le[1,2], Phuc Nguyen[1,2,*]

[1]Faculty of Information Systems, University of Economics and Law, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{thuda22411c, anhttq22411c, vydt22411c, tuttc22411c}@st.uel.edu.vn, {sulh, phucnq}@uel.edu.vn
*Corresponding Author: phucnq@uel.edu.vn

Received January 6, 2025, revised March 9, 2025, accepted March 11, 2025.

Abstract. *Classifying student emails is a crucial aspect of effective information management in higher education. This study builds upon previous research by the team, which demonstrated that combining deep learning architectures with keyword extraction techniques achieved an accuracy of 93%, surpassing traditional deep learning models by 1%* [1]. *In the current study, an enhanced methodology is introduced, incorporating email summarization to refine classification performance. While the overall accuracy remains at 93%, the approach significantly improves the accuracy across individual labels, addressing imbalances in classification outcomes. These findings emphasize the value of combining feature extraction and summarization techniques in managing complex email datasets. The proposed framework not only enhances email classification in academic contexts but also has broader implications for text classification, information management, and big data analytics, showcasing its adaptability and potential impact.*

Keywords: Email Categorization, Semantic Analysis, Text Summarization, Deep Learning for Text Classification, Data Labeling Techniques.

1. **Introduction.** Computer-mediated communication (CMC) has long been acknowledged as a key element in higher education [2]. Over the years, the role of CMC, particularly through email, has evolved beyond mere academic utility to become an essential element of professional communication. This shift underscores the increasing reliance on digital platforms for both educational and workplace interactions [3, 4]. Despite the numerous advantages that CMC offers, such as enhanced connectivity and streamlined information exchange, it has also given rise to challenges, particularly the phenomenon known as "email overload". First identified by Whittaker and Sidner [5], this issue refers to the overwhelming influx of messages that complicates effective management and prioritization of communications. Extensive research has demonstrated that this overload can detrimentally impact workplace productivity and individual well-being. Studies indicate that excessive email volume can lead to burnout and reduced work engagement [6], while also affecting students' mental health by contributing to stress and anxiety, ultimately decreasing their productivity [7]. These findings highlight an urgent need to address the challenges posed by email overload. Developing effective systems for email classification and management is crucial to mitigate its adverse effects.

Email classification, which falls under the broader category of text classification, involves using key attributes such as sender information, recipient details, subject lines,

and content analysis to identify the primary topics of emails and categorize them accordingly. However, this process is complicated by the vast diversity of email datasets encountered in practice. To tackle these complexities, this study employs advanced deep learning architectures known for their ability to uncover intricate patterns within data. By leveraging deep learning architecture, it aims to enhance classification accuracy and performance compared to traditional machine learning architecture approaches [8, 9]. Specifically, this research integrates feature extraction and email summarization within a deep learning framework. By summarizing email content and extracting salient features, it constructs a comprehensive feature space that enables models to differentiate between similar emails with subtle variations in meaning. Furthermore, this approach effectively addresses challenges associated with large datasets and diverse types of emails, facilitating improved generalization across various contexts. By harnessing the power of deep learning to reveal complex patterns, this method ensures robust classification capabilities even for nuanced and context-dependent email categories.

The objective of this research is to propose an effective solution that enables automatic email classification systems to operate accurately and efficiently, thereby improving the email processing workflow within academic environments. This approach can also be extended to other domains such as text classification in economics, social sciences, and customer services, where managing and processing large datasets is crucial. The paper is structured as follows: Section II reviews related studies. Section III presents the proposed approach. Section IV describes the dataset, results, and performance analysis. Section V concludes and outlines directions for future research.

2. **Related works.** Text classification has long been a focus of research, particularly in the context of Vietnamese language processing. This field has seen diverse approaches yielding valuable contributions. Nguyen Phuoc Vinh and Ha Hoang Kha [10] introduce a benchmark Vietnamese online news dataset designed for multi-label classification, comprising 30 topics labeled similarly to editorial practices. By modifying the traditional text classification pipeline, the study reduces computational complexity through direct dimensionality reduction in the TF-IDF weighting step, omitting feature selection algorithms. Despite weaker feature vectors, the use of neural network models achieves competitive or slightly improved performance. Future directions include extending the dataset with additional classes and exploring enhanced feature extraction techniques to improve classification accuracy. Dang Van Thin et al. [11] proposed a multi-task deep learning model to simultaneously address aspect detection and sentiment polarity classification at the document level, applied to Vietnamese datasets. Experimental results on two VLSP benchmark datasets (restaurant and hotel domains) demonstrate the model's superior performance, achieving state-of-the-art F1_scores of 64.78% and 70.90%, respectively. Strengths include the effective utilization of pre-trained embeddings, such as word2vec and Multi-embedding, though limitations remain in generalizability across other domains. Future work will focus on detailed result analysis and evaluating the effectiveness of new pre-trained embeddings with the proposed model. Meanwhile, Nguyen et al. [12] introduced the Social Media Text Classification Evaluation (SMTCE) benchmark for Vietnamese, focusing on four tasks: emotion recognition, constructive speech detection, hate speech detection, and complaint comment detection. The authors evaluated multilingual (mBERT, XLM-R) and monolingual BERT-based models (PhoBERT, viBERT), with monolingual models achieving state-of-the-art performances across all tasks, including F1_Scores of 65.44% (VSMEC) and 95.26% (ViOCD). However, some of the models only showed slight improvements over the baseline results. While ensemble methods achieved better performance, they came at the cost of slower processing times. This highlights the trade-off

between performance gain and computational efficiency in choosing the most suitable model for the task at hand. In[13], the authors advanced sentiment classification by combining PhoBERT with CNN, achieving superior performance compared to conventional approaches such as SVM and XGBoost, with a precision of 0.9405, a recall of 0.8823, and an F1 score of 0.9104. Their approach effectively captures local features but faces challenges with training time. These studies highlight the critical role of specialized models and dataset-specific strategies in achieving high-performance text classification, paving the way for further advancements in preprocessing and representation optimization.

Beyond Vietnamese text classification, general text classification methods, especially those using BERT, have garnered significant attention from international researchers, addressing diverse challenges. For example, Zhengjie Gao et al. [14] explored the application of BERT in target-dependent sentiment classification, a sub-task of aspect-based sentiment analysis (ABSA), and demonstrated that BERT significantly raises the baseline for text representation in this domain. While traditional feature engineering with classifiers outperforms embedding-based models, BERT achieves state-of-the-art results on SemEval-2014 and a Twitter dataset through its context-aware representation and target-focused modifications. However, improvements over the vanilla BERT implementation are modest, suggesting that either the potential of BERT is not fully exploited or it has already established a high-performance ceiling. Challenges remain in accurately classifying neutral or mixed-polarity cases, necessitating more training data and advanced analysis methods for further improvement. Future research will focus on these challenges and the exploration of better network structures. According to the work by Cao et al. [15], the challenge of detecting bilingual and multi-type spam was addressed by proposing a novel model based on Google's multilingual BERT (M-BERT) and utilizing a newly developed bilingual multi-type spam dataset. The model incorporates optical character recognition (OCR) to extract text from image-based spam. Experimental results reveal the model achieves an accuracy of 96.48%, surpassing traditional approaches, with an efficient training time of 0.3168 seconds per step. This demonstrates the model's effectiveness and practicality in detecting diverse spam messages across multiple languages and types, contributing a significant advancement in spam detection technology. Similarly, Rasmy et al. [16] introduced Med-BERT, a contextualized embedding model adapted from the BERT framework to structured electronic health record (EHR) data, addressing the challenge of limited training data in deep learning-based predictive healthcare models. Pretrained on EHRs of 28,490,650 patients, Med-BERT demonstrates significant improvements in disease prediction tasks, with AUC increases of 1.21–6.14% and over 20% in small dataset scenarios compared to traditional models. This model offers potential cost reductions in data collection and accelerates AI adoption in healthcare, though computational resource requirements and ethical considerations for data use highlight areas for careful attention in future implementations. BERT's ability to capture deep contextual understanding has also been highlighted by Kamal Taha et al. [17], who emphasized its effectiveness in sentiment analysis. Despite its strengths, BERT's reliance on high-quality data and significant computational resources remains a challenge. Future research could explore optimizing BERT's pretraining process and extending its applications to new domains, promoting both efficiency and accessibility in text classification advancements.

Optimizing text classification with BERT can be significantly enhanced by integrating text summarization techniques. Condensing content while maintaining key information boosts processing speed and classification accuracy. Various studies provide valuable insights for improving model architectures. Reda Elbarougyet al. [18] proposed a method for summarizing Arabic texts using the Al-Khalil morphological analyzer to address the challenges of noun extraction, which is difficult due to the complex grammatical structure

of the Arabic language. This approach constructs a graph-based system where sentences are represented as nodes, and cosine similarity between sentences is used to determine edge weights. The results indicate that the Modified PageRank algorithm, applied with 10,000 iterations, achieves higher effectiveness compared to other Arabic text summarization techniques, with an F-measure score of 67.98. However, the study suggests improvements in practical applicability and recommends expanding the testing scale to enhance the accuracy and reliability of the generated summaries. Liu et al. [19] introduced BERTSUM, a variant of BERT, which, in terms of extractive summarization, outperforms the previous best system by 1.65 points in ROUGE-L scores on the CNN/DailyMail dataset. Extractive summarization is highlighted as a crucial task in natural language processing, with applications in generating concise document versions while preserving essential information. Various configurations of BERT for this task are explored, revealing that a flat architecture with inter-sentence Transformer layers delivers the best performance. Despite recent advancements in neural models for extractive summarization, improvements in automatic evaluation metrics like ROUGE have reached a plateau, indicating that BERT's pre-training on extensive datasets can significantly enhance summarization capabilities. BARTpho, introduced by Nguyen Luong Tran et al. [20], includes two pre-trained monolingual models, BARTphosyllable and BARTphoword, optimized for generative NLP tasks in Vietnamese. It surpasses mBART in Vietnamese text summarization and excels in capitalization and punctuation restoration. Evaluations show BARTpho's superiority, making it a state-of-the-art solution. In the study by Thang Le Ngoc et al. [21], they proposed a model for Vietnamese online newspaper summarization using the LexRank algorithm, incorporating Tags words and named entities for better sentence similarity calculation. The model outperforms basic LexRank in accuracy. Future work will focus on optimizing similarity calculations and addressing redundancy to further enhance efficiency and performance.

Previous research in text classification and summarization underscores the importance of specialized models and strategies for domain-specific tasks, highlighting advancements and challenges such as balancing performance with computational efficiency. Integrating text summarization methods with classification techniques has shown promise for improving accuracy and scalability. Building on this foundation, our study aims to enhance student email classification by combining deep learning architecture with feature extraction and summarization techniques. This approach not only refines email categorization but also lays the groundwork for broader applications across diverse text classification domains.

3. **Proposed framework.**

The classification of Vietnamese educational emails requires addressing specific linguistic and contextual challenges. Our previous work with PhoBERT-Based revealed potential improvements in semantic representation and feature enrichment [1]. This research presents an enhanced framework incorporating text summarization while maintaining the core classification approach. The framework consists of four stages: *i) information extraction and preprocessing, ii) feature extraction, iii) training the classification model, iv) prediction and model evaluation.*

The information extraction and preprocessing stage structures raw email data across three components: sender address, title, and content. This phase encompasses character encoding normalization for Vietnamese diacritics, removal of special characters and redundant whitespace, and elimination of standard email signatures while preserving semantic markers. These operations produce normalized text suitable for subsequent processing.

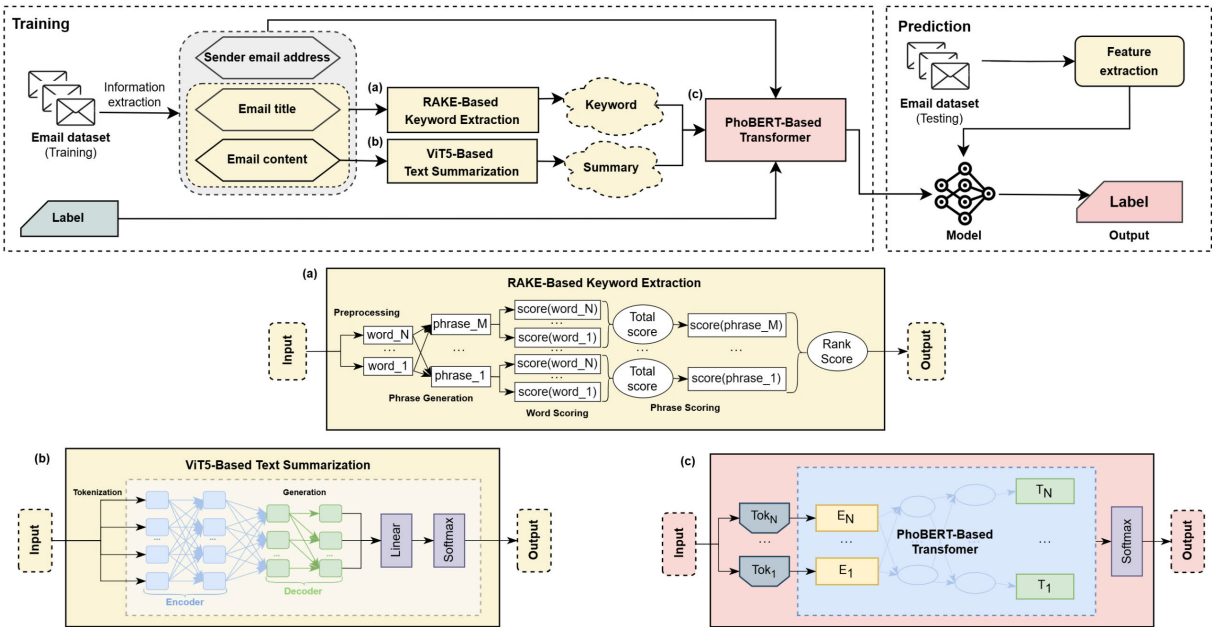

FIGURE 1. Overview of the proposed email classification framework. (a) RAKE-Based Keyword Extraction process. (b) ViT5-Based Text Summarization process. (c) The architecture of PhoBERT-Based Transformer model.

Our framework's key innovation lies in the enhanced feature extraction stage, which integrates keyword extraction with text summarization. This dual-feature approach represents a significant advancement over our previous work, which solely relied on keyword extraction. By incorporating text summarization, we achieve a more comprehensive semantic representation of email content, capturing both specific terminological elements and broader contextual information.

The keyword extraction implements RAKE (Rapid Automatic Keyword Extraction), selected for its efficiency in identifying salient terms within email titles and content. As shown in Figure 1(a), the process begins with text preprocessing and stop-word removal, followed by contextually-aware phrase generation. The algorithm calculates word significance through frequency and distinctiveness metrics, producing phrase scores based on word weights and position. The system selects the top 10 scoring phrases per label as keywords.

The text summarization component utilizes the ViT5-Based model for generating compact representations of email content, as illustrated in Figure 1(b). The process begins with a specialized tokenizer that transforms input text into encoded hidden states. The model's decoder employs beam search with a beam size of 4, enabling comprehensive evaluation of multiple potential sequences at each generation step. To maintain optimal summary quality, we implement a length penalty coefficient of 2.0, ensuring summaries remain between 40 and 600 tokens—striking an ideal balance between completeness and conciseness. The system incorporates early stopping mechanisms to prevent unnecessary token generation once optimal sequences are identified. Finally, a linear softmax transformation converts the generated tokens into coherent text summaries.

The training stage implements PhoBERT-Based Transformer architecture for pattern learning. The system encodes extracted features, keywords, and summaries into input tokens $(Tok_1, \ldots, Tok_N)$, converting them to embeddings $(E_1, \ldots, E_N)$. These pass through

transformer layers, generating semantic vectors *(T_1,..., T_N)*. The final layer applies Softmax for classification probabilities.

The evaluation stage measures framework performance using a test dataset, comparing predicted labels with ground truth values to assess classification accuracy.

The framework presents a structured approach to email classification by integrating feature extraction techniques with deep learning methodologies. It leverages PhoBERT's language processing capabilities for Vietnamese text analysis and employs a dual feature extraction methodology. The framework implements feature representation that incorporates both specific terminological elements and broader content context, establishing a methodological foundation for future empirical evaluation.

4. **Experiments.**

4.1. **Dataset.** This study involves the development of a dataset comprising student emails from universities and colleges nationwide. Each email includes essential attributes such as sender, recipient, subject, and content, categorized into five labels: "HOC BONG", "HOC TAP", "CHUONG TRINH TRAO DOI - DU HOC", "CAC KHOAN PHI", and "KHAO SAT", detailed in Table 1.

Table 1. Distribution of email categories in the dataset

Label (Vietnamese)	Description	No. email
"HOC TAP"	Emails regarding tasks, due dates, instructions, enrollment, timetable updates, learning resources, classes, and extra meetings.	1001
"KHAO SAT"	Messages inviting students to participate in surveys, provide feedback, or contribute to evaluations aimed at gathering opinions or data for research purposes.	720
"HOC BONG"	Announcements and updates on scholarship opportunities, detailing requirements, application processes, types of scholarships, and associated events or programs.	749
"CAC KHOAN PHI"	Emails addressing payment obligations, such as tuition fees, administrative charges, or insurance premiums, often including deadlines and payment instructions.	735
"CHUONG TRINH TRAO DOI - DU HOC"	Communications highlighting opportunities for exchange programs or studying abroad, covering eligibility, application steps, and relevant informational sessions.	710

The "Label (Vietnamese)" column in our dataset defines the classification categories, while the "Description" column elaborates on each label, offering a detailed explanation of the criteria used for classification. The "No. email" column shows the number of emails associated with each label throughout the dataset.

The dataset is divided into three subsets: 70% for training, 15% for testing, and 15% for validation. This distribution ensures balanced representation across all categories, a crucial factor for developing an effective classification model. The training set provides the algorithm with a substantial amount of labeled data, enabling it to learn and identify

patterns within each category. This diverse training data enhances the model's generalization ability to new, unseen data. The testing set serves as an independent evaluation benchmark, assessing the model's performance on previously unseen emails. Finally, the validation set facilitates iterative model tuning by allowing for adjustments to parameters and hyperparameters based on performance metrics.

To effectively analyze student institution communication, we categorized emails into five distinct labels, carefully selected to reflect the most frequent communication themes observed during data collection. This approach ensures that each category accurately represents the range of student inquiries and interactions, thereby enhancing the practical value of our classification system. Furthermore, we opted for a single-label classification approach to enhance classification accuracy and maintain clarity, excluding multi-label emails to avoid complexity and ambiguity. This decision simplifies the task for both the model and users by assigning a single label based on the email's primary topic, allowing for clearer distinctions between categories and reducing potential confusion from overlapping topics.

Additionally, it is important to note that our dataset is presented in Vietnamese, specifically catering to the linguistic context and cultural nuances relevant to our research focus. This choice not only ensures that our findings are applicable within Vietnamese educational settings but also addresses potential challenges associated with language processing in non English contexts. By maintaining a balanced representation across labels while using a single-label classification strategy, we effectively reduce complexity and address potential imbalances in email distribution.

Overall, this comprehensive approach supports effective training and evaluation processes while enhancing performance in classification tasks. By carefully structuring our dataset in this manner, we aim to contribute valuable insights into student communications within higher education and provide a foundation for future research in this domain.

4.2. **Implementation Details.** This research leveraged the Google Colab platform with Google Compute Engine (TPU) for its robust and scalable environment, facilitating the execution of computationally demanding deep learning tasks. The primary framework employed was PyTorch, renowned for its flexibility and dynamic computational graph capabilities. The Hugging Face Transformers library was integrated to access pre-trained models and advanced NLP implementations.

The PhoBERT-Based Transformer model was configured with a batch size of 16, a learning rate of $2e^{-5}$, and 12 hidden layers with 768 features, adhering to standard transformer architectures. A batch size of 16 ensures a balance between computational efficiency and the stability of gradient updates, making it suitable for fine-tuning on moderate-sized datasets. The learning rate of $2e^{-5}$, commonly used in fine-tuning pre-trained transformers, provides stable convergence and prevents drastic weight updates that could destabilize training. Training was limited to 4 epochs to avoid overfitting, ensuring the model generalizes well to unseen data. The output layer utilizes Softmax activation, which is crucial for generating a multi-class probability distribution in classification tasks. AdamW optimization was employed to further enhance generalization by decoupling weight decay from the gradient update process. This separation ensures that weight regularization does not interfere with the learning dynamics, thereby reducing overfitting and promoting stable and efficient training, especially in large-scale transformer models [22].

This experimental setup was carefully designed to optimize the PhoBERT model's performance for the specific challenges of student email categorization within the higher education domain.

In this study, four experimental methods were implemented to optimize the PhoBERT model's performance in addressing the unique challenges of categorizing student emails within the higher education context. These methods are as follows:

- (PB) - PhoBERT-Based: This method utilizes the standard deep learning architecture to classify Vietnamese text, serving as the foundational baseline for evaluating student email classification.
- (PBK) - PhoBERT-Based with Keyword Extraction: The standard model is enhanced by incorporating keyword extraction, which improves the model's ability to identify critical elements within email content.
- (PBS) - PhoBERT-Based with Text Summarization: This approach combines the deep learning architecture with text summarization, enabling the model to condense essential information, filter out noise, and achieve more accurate classifications.
- (PBKS) - PhoBERT-Based with Keyword Extraction and Text Summarization: This method integrates both keyword extraction and text summarization into the standard model, leveraging the strengths of both techniques to optimize classification performance.

The detailed outcomes of these methods are presented in Table 2, providing insights into the effectiveness of each approach for this classification task.

4.3. **Results and Evaluation.** To enhance classification accuracy and mitigate overfitting, we employed K-Fold Cross-Validation with five folds. This robust technique ensures model generalizability by iteratively training and evaluating the model on different subsets of the data. Each fold serves as the test set once, maximizing data utilization and providing a comprehensive assessment of model performance across various data segments. The model was optimized over four epochs during each training session, allowing for effective learning from the data.The final performance metrics, derived from Table 2, include True

Table 2. The classification results for each label

Label (Vietnamese)	Precision				Recall				F1_Score			
	PB	PBK	PBS	PBKS	PB	PBK	PBS	PBKS	PB	PBK	PBS	PBKS
"CHUONG TRINH TRAO DOI - DU HOC"	0.91	0.94	**0.96**	**0.96**	0.90	**0.91**	0.89	0.89	0.90	**0.92**	**0.92**	**0.92**
"HOC BONG"	0.97	0.97	**0.98**	**0.98**	0.93	0.93	**0.94**	**0.94**	0.95	0.95	**0.96**	**0.96**
"HOC TAP"	0.86	0.88	**0.89**	**0.89**	0.97	0.97	**0.99**	**0.99**	0.91	0.92	**0.93**	**0.93**
"CAC KHOAN PHI"	**0.93**	**0.93**	0.91	0.91	0.93	**0.95**	**0.95**	**0.95**	0.93	**0.94**	0.93	0.93
"KHAO SAT"	**1.00**	0.97	0.96	0.96	0.86	0.86	**0.87**	**0.87**	0.93	0.91	0.91	0.91
Accuracy									0.92	**0.93**	**0.93**	**0.93**
Macro avg	0.93	**0.94**	**0.94**	**0.94**	0.92	**0.93**	**0.93**	**0.93**	0.92	**0.93**	**0.93**	**0.93**
Weighted avg	0.93	0.93	**0.94**	**0.94**	0.92	**0.93**	**0.93**	**0.93**	0.92	**0.93**	**0.93**	**0.93**

Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN), which are crucial for evaluating model performance, are listed as follows:

- *TP (True Positive)*: The number of positive samples correctly identified.
- *FP (False Positive)*: The number of negative samples incorrectly labeled as positive.
- *FN (False Negative)*: The number of positive samples incorrectly labeled as negative.
- *TN (True Negative)*: The number of negative samples correctly identified.

To quantify the model's classification performance, several standard metrics were calculated, as detailed below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision \times Recall} \tag{4}$$

The experimental results in Table 2 highlight the significant advancements in email classification achieved by the proposed architecture, particularly for complex and semantically challenging categories. While maintaining an overall accuracy of 93%, the model demonstrates notable improvements in individual label performance. Categories like "HOC BONG" and "HOC TAP" benefit greatly from the architecture's ability to capture critical features and contextual nuances, resulting in reduced classification errors and improved precision.

Building on previous methods that primarily focused on keyword extraction, the proposed PBS approach achieves a more balanced and robust classification performance, particularly in challenging scenarios. By integrating text summarization, the model effectively captures essential information, reduces noise, and enhances its ability to generalize across diverse and complex datasets. This synergy between techniques lays the foundation for the model's superior classification accuracy.

Experimental results further reveal that combining keyword extraction and text summarization - PBKS yields similar outcomes to using text summarization alone. This underscores the distinct strengths of each method: text summarization excels at condensing critical information, while keyword extraction improves the detection of key features. Additionally, incorporating advanced deep learning architectures with these techniques highlights the significant potential for improving classification performance, especially in more complex contexts.

Categories such as "HOC BONG" stand out as the proposed architecture excels in handling long emails where critical keywords appear infrequently—a challenge that traditional keyword extraction methods struggled to address accurately. By leveraging text summarization, the model focuses on essential features, enabling precise classification even in complex contextual scenarios. Similarly, the "HOC TAP" category, characterized by its high contextual complexity, benefits significantly from the summarization process, which allows the model to efficiently identify key patterns and features without relying on full-text processing. In contrast, the performance of the "KHAO SAT" label is less consistent, likely due to the wide variation in email content within this category. While the content is diverse, the dataset's limited size may not provide sufficient examples for the model to learn robust classification rules, leading to reduced accuracy compared to other

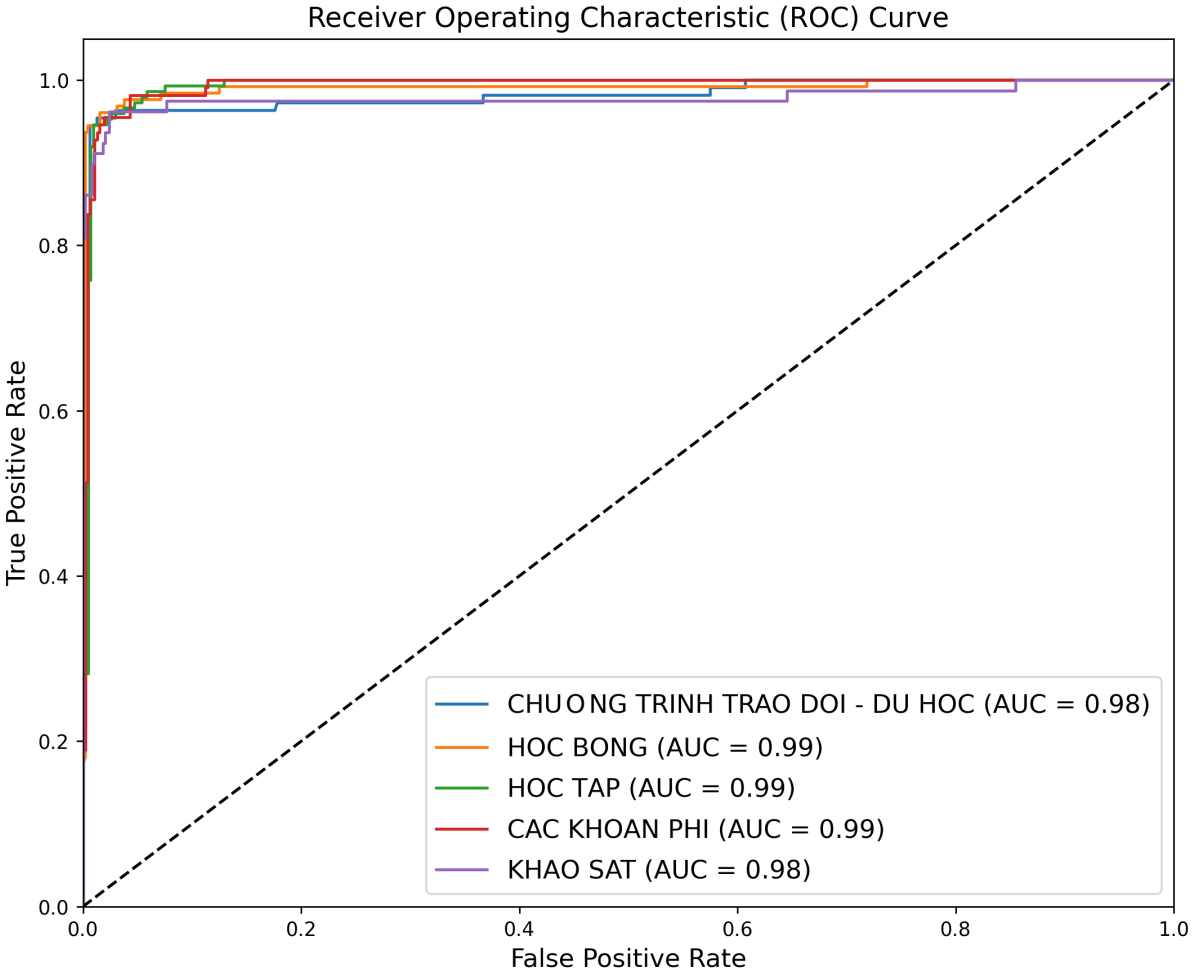

FIGURE 2. Our experimental results with PBKS are presented through the AUC and ROC curves.

categories. Nonetheless, the proposed architecture demonstrates significant improvements over earlier methods, minimizing classification errors and delivering reliable results even in challenging scenarios.

Figure 2 presents the AUC and ROC curves of the proposed PBKS method, illustrating consistent and reliable performance in email classification. These metrics affirm the model's ability to maintain high classification accuracy, even when addressing complex or imbalanced categories. The integration of keyword extraction and text summarization techniques not only aids the model in accurately identifying semantic features but also plays a crucial role in minimizing misclassifications for individual labels. This demonstrates the method's potential for achieving precise and consistent classification across diverse scenarios.

The proposed architecture demonstrates balanced and reliable classification performance, particularly for complex or imbalanced categories. While a significant advancement in email classification, limitations in computational efficiency and scalability may hinder real-time deployment in dynamic, high-demand environments. This consistently high performance, surpassing standard deep learning architectures, underscores the model's ability to accurately distinguish between classes. The integration of advanced techniques, such as keyword extraction and text summarization, enhances its discriminatory power by facilitating precise feature recognition and reducing misclassifications. The model's

robust performance across diverse email structures and content complexities highlights its reliability and scalability for practical applications in email classification tasks.

This methodology demonstrates significant potential for broader application across diverse domains, including text data analysis, natural language processing, customer service automation, finance, and healthcare communication systems. While promising, enhancing computational efficiency and real-time deployment capabilities is essential for scaling its application within complex and demanding environments. Overcoming these challenges will be paramount in optimizing operational effectiveness and ensuring the classification system's adaptability to evolving user needs in dynamic contexts.

This consistently high performance highlights not only the model's ability to accurately distinguish between classes but also its superiority over standard deep learning architectures. The integration of advanced techniques such as keyword extraction and text summarization enhances its discriminatory power, facilitating precise feature recognition while reducing misclassifications. The model's adeptness at handling varying email structures and content complexities is evident in its robust performance, showing its reliability and scalability for practical applications in email classification tasks.

Moreover, this methodology holds promise for broader applications beyond email classification, extending into fields such as text data analysis, natural language processing, customer service automation, finance, and healthcare communication systems. Nevertheless, there remains a need for improvements in computational efficiency and real-time deployment capabilities to enable scaling within more complex and high-demand environments. Addressing these aspects will be crucial for maximizing operational effectiveness and ensuring that this advanced classification system can meet evolving user needs in dynamic contexts.

5. **Conclusions.** This study proposes a novel framework for student email classification, leveraging the integration of keyword extraction and text summarization techniques within the deep learning architecture. This approach demonstrates significant potential for improving performance across specific labels by effectively capturing semantic features. However, the current implementation encounters challenges in real-time application, stemming from limited computational efficiency, as well as the small size and high variability of the dataset. Future research will focus on model compression, optimization for large-scale deployment, and dataset expansion to include English-language emails. These advancements aim to enhance multi-label generalizability and broaden the framework's applicability across diverse educational and professional domains.

Acknowledgment. This research is funded by University of Economics and Law, Vietnam National University Ho Chi Minh City / VNU-HCM.

REFERENCES

[1] Duong, T., Tran, A., Dang, V., Tran, T., Le, H. S. & Nguyen, P. (2024, December). Enhancing Educational Email Management Efficiency using Deep Learning. In *2024 21st International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE.

[2] Steeples, C., Goodyear, P., & Mellar, H. (1994). Flexible learning in higher education: the use of computer-mediated communications. *Computer & Education*, 22, 83-90.

[3] Derks, D., & Bakker, A. B. (2010). The Impact of E-mail Communication on Organizational Life. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 4(1), Article 4.

[4] Uddin, S., & Jacobson, M. J. (2013). Dynamics of email communications among university students throughout a semester. *Computers & Education*, 64, 95–103.

[5] Whittaker, S., & Sidner, C.L. (1996). Email overload: exploring personal information management of email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[6] Reinke, K., & Chamorro-Premuzic, T. (2014). When email use gets out of control: Understanding the relationship between personality and email overload and their impact on burnout and work engagement. *Computers in Human Behavior*, 36, 502–509.
[7] Wood, K. E., & Krasowski, M. D. (2020). Academic E-Mail overload and the burden of "Academic spam". *Academic Pathology*, 7, 2374289519898858.
[8] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: a comprehensive review. In *ACM computing surveys (CSUR)*), 54(3), 1-40.
[9] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
[10] Vinh, T. N. P., & Kha, H. H. (2021). Vietnamese news articles classification using neural networks. In *Journal of Advances in Information Technology*, 12(4).
[11] Van Thin, D., Nguyen, D., Van Nguyen, K., Nguyen, N. L., & Nguyen, A. H. (2020). Multi-task learning for aspect and polarity recognition on Vietnamese datasets. In *Communications in computer and information science*, (pp. 169–180).
[12] Nguyen, L., Nguyen, K.V., & Nguyen, N.L. (2022). SMTCE: A Social Media Text Classification Evaluation Benchmark and BERTology Models for Vietnamese. In *Pacific Asia Conference on Language, Information and Computation.*
[13] Loc, C. V., Viet, T. X., Viet, T. H., Thao, L. H., & Viet, N. H. (2023). Pre-Trained Language Model-Based Deep Learning for sentiment classification of Vietnamese feedback. In *International Journal of Computational Intelligence and Application*, 22(03).
[14] Gao, Z., Feng, A., Song, X., & Wu, X. (2019). Target-Dependent Sentiment Classification with BERT. *IEEE Access*, 7, 154290–154299.
[15] Cao, J., & Lai, C. (2020). A bilingual multi-type spam detection model based on M-BERT. In *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 1–6.
[16] Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1), 86.
[17] Taha, K., Yoo, P. D., Yeun, C., Homouz, D., & Taha, A. (2024). A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. *Computer Science Review*, 54, 100664.
[18] Elbarougy, R., Behery, G., & Khatib, A. E. (2019). Extractive Arabic text summarization using modified PageRank algorithm. *Egyptian Informatics Journal*, 21(2), 73–81.
[19] Liu, Y. (2019). Fine-tune BERT for extractive summarization. *Computation and Language.*
[20] Tran, N. L., Le, D., & Nguyen, D. Q. (2022). BARTPho: Pre-trained Sequence-to-Sequence models for Vietnamese. *Interspeech 2022.*
[21] Ngoc, T. L., & Minh, L. Q. (2023). Vietnamese online newspapers summarization using LexRank. In *Development of Social and Humanitarian Knowledge: New Directions and Prospects.*
[22] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR).*