

ابوبكر حسين

DistilBERT for Email Classification and Priority.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:400682990

Submission Date

Jul 22, 2026, 10:15 PM GMT+3

Download Date

Jul 27, 2026, 10:23 PM GMT+3

File Name

DistilBERT for Email Classification and Priority.docx

File Size

2.0 MB

19 Pages

4,180 Words

25,031 Characters

18% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

-  **62 Not Cited or Quoted 17%**
Matches with neither in-text citation nor quotation marks
-  **6 Missing Quotations 1%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 11%  Internet sources
- 11%  Publications
- 13%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 62 Not Cited or Quoted 17%**
Matches with neither in-text citation nor quotation marks
- 6 Missing Quotations 1%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 11% Internet sources
- 11% Publications
- 13% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Publication	Victor Jima. "Integrating Traditional SMEs into Hybrid Digitalized Trade Navigatio...	1%
2	Submitted works	Vrije Universiteit Amsterdam on 2024-11-26	1%
3	Internet	assets.pubpub.org	<1%
4	Internet	www.uyik.org	<1%
5	Submitted works	University of Northumbria at Newcastle on 2024-08-28	<1%
6	Internet	ojs.bonviewpress.com	<1%
7	Internet	engrxiv.org	<1%
8	Publication	"Smart Data Intelligence", Springer Science and Business Media LLC, 2024	<1%
9	Internet	www.ijraset.com	<1%
10	Internet	www.jihmsp.org	<1%

11	Internet	ris.utwente.nl	<1%
12	Publication	"AI Technologies for Information Systems and Management Science", Springer Sc...	<1%
13	Submitted works	University of Mpumalanga on 2025-11-14	<1%
14	Submitted works	Midocean University on 2025-09-05	<1%
15	Submitted works	Southern Cross University on 2024-05-28	<1%
16	Internet	www.mdpi.com	<1%
17	Internet	pure.nwpu.edu.cn	<1%
18	Publication	Rafael da Fonseca Lameiro. "Análise de dados e aprendizado de máquina para o d...	<1%
19	Submitted works	Tilburg University on 2025-01-10	<1%
20	Submitted works	University of Westminster on 2026-04-12	<1%
21	Publication	Bhabani Prasanna Pattanaik, Dharendra Kumar Shukla, Sambit Satpathy, Kamala...	<1%
22	Submitted works	Indian Institute of Technology, Madras on 2026-03-09	<1%
23	Submitted works	Islington College,Nepal on 2026-01-20	<1%
24	Publication	S. Prasad Jones Christydass, Nurhayati Nurhayati, S. Kannadhasan. "Hybrid and A...	<1%

25	Internet	github.com	<1%
26	Internet	link.springer.com	<1%
27	Internet	www.irrodl.org	<1%
28	Submitted works	George Mason University on 2025-11-23	<1%
29	Internet	jurnal.polibatam.ac.id	<1%
30	Internet	www.kaggie.com	<1%
31	Submitted works	Kaplan College on 2025-03-20	<1%
32	Publication	Poonam Nandal, Tapas Kumar, Meeta Singh, Mamta Dahiya. "Progressive Compu...	<1%
33	Internet	www.nature.com	<1%
34	Submitted works	Innopolis University on 2020-05-12	<1%
35	Submitted works	JB Institute of Engineering And Technology on 2026-04-08	<1%
36	Submitted works	Liverpool John Moores University on 2023-02-28	<1%
37	Submitted works	Midocean University on 2025-09-05	<1%
38	Submitted works	Universiti Teknologi MARA on 2026-07-21	<1%

39	Internet	papers.academic-conferences.org	<1%
40	Internet	research.aalto.fi	<1%
41	Internet	www.atlantis-press.com	<1%
42	Publication	A. Punitha, S. Syedakbar, S. Jeyasudha. "Intelligent and Sustainable Systems: AI, G...	<1%
43	Submitted works	Federal University of Technology on 2023-12-28	<1%
44	Publication	Hsu, Kai-Hsun. "Machine Learning-Based Tool Wear Estimation for Milling Proces...	<1%
45	Submitted works	University of East London on 2026-05-08	<1%
46	Submitted works	University of Northumbria at Newcastle on 2023-07-13	<1%
47	Submitted works	University of Westminster on 2021-05-02	<1%
48	Submitted works	Indian Institute of Technology Patna on 2022-08-31	<1%

A Multi-Task DistilBERT Framework for Multi-Label Email Classification and Priority Prediction

Researcher:

Abubaker Hussein (ORCID: 0009-0009-4342-5862)

Artificial Intelligence, College of Informatics, Midocean University, UAE.

Supervisor:

Dr. Amjad Qatish

College of Informatics, Midocean University, Comoros

Email: amjadqtaish@midocean.edu.km

Abstract

Email management is getting more difficult as organizations receive increasing amounts of complex communications. Current approaches to email categorization mainly focus on individual tasks, such as spam detection or single-label categorization, making them ineffective at processing emails that involve multiple topics and varying priorities. In this paper, we propose a multi-task learning approach with a DistilBERT encoder for concurrent multi-label email classification and priority prediction. Our proposed architecture uses the common encoder along with two separate task-specific decoders. The model was trained on a selected and labeled sample of the Enron Email Dataset by applying a stratification technique. Experimental results show that the DistilBERT model achieved a Macro F1 score of 0.9672 and Hamming Loss of 0.0175 for category classification and 95.29% accuracy with a Macro F1 score of 0.9327 for priority prediction. Moreover, thresholding and LIME helped improve performance and provide greater model interpretability.

Keywords: Email Classification, Multi-Label Classification, Email Priority Prediction, DistilBERT, Multi-Task Learning, Explainable Artificial Intelligence, LIME, Natural Language Processing.

1. Introduction

Despite advancements in technological means of communication, email is still one of the most prevalent forms of communication used in governmental, educational, and corporate bodies. As the number of incoming emails keeps growing, organizing emails based on their relevance and urgency has become quite tedious, resulting in delayed actions and inefficiencies (Khare et al., 2022).

In previous studies, email categorization has been largely concentrated on the problem of spam email identification or single-label email classification, utilizing conventional machine learning methods like Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Naïve Bayes (Iqbal & Khan, 2022; Fareed et al., 2025). While yielding satisfactory classification results, these techniques mostly use statistical models to represent email data and thus lack the capability of semantic understanding of email texts.

The latest research on Transformer-based language models has shown that there is high efficiency of contextual text representation for natural language processing tasks. Models that are based on BERT have shown better results in many classification tasks of texts since they learn the semantic relations between the words instead of word frequencies (Duong et al., 2025; Wang et al., 2024). However, the majority of current research papers still deal with only one task, such as e-mail classification or urgency estimation.

Other limitations can be mentioned regarding the literature review. Many researchers have studied spam or phishing detection using the binary classification task (Jáñez-Martino et al., 2022; Salloum et al., 2022; Ahmed et al., 2022), while other studies focused only on the single-label email categorization (Bhalla et al., 2021; Fareed et al., 2025). Moreover, current approaches to priority prediction do not consider the multi-topic nature of the email (Chuttur et al., 2023; Hirpara & Polara, 2025).

From the above observations, it can be seen that there is indeed an identified gap related to the development of an intelligent email management system capable of conducting multiple label category classifications and priority prediction at the same time, while at the same time being interpretable.

To fill the above research gap, the current study designs a multi-task learning architecture based on DistilBERT. In this multi-task learning framework, multiple email categories as well as the corresponding priority level of each email are predicted together through a common encoder as well as two distinct output layers for each of the tasks. Moreover, threshold tuning will be used to enhance the multi-label category prediction performance, and Local Interpretable Model-agnostic Explanations (LIME) will be applied to ensure interpretation of the predictions.

In the current study, the proposed approach is evaluated using the Enron Email dataset. It is compared against conventional machine learning approaches, where Logistic Regression is used for category classification, and Linear SVM is used for priority prediction.

2. Related Work

Extensive research work has been carried out to address the problem of email classification through the use of traditional and deep machine learning algorithms. In early studies, efforts have mostly concentrated on spam detection and binary classification using traditional approaches for feature extraction, such as TF-IDF, in conjunction with Logistic Regression, Support Vector Machines (SVM), Decision Tree, and Random Forest classifiers (Iqbal & Khan, 2022; Ahmed et al., 2022). Though high accuracy was attained through these approaches using benchmark datasets, they were not able to leverage semantic relationships between words within the text as they utilized hand-crafted features.

Recently, efforts have increasingly turned towards transformer-based models in order to improve email understanding capabilities. For instance, Duong et al. (2025) have developed the AEEM (educational email management) framework using a PhoBERT-based model and have attained about 93% classification accuracy. Additionally, Hirpara and Polara (2025) have come up with the InboxG framework, which uses AI for email management tasks such as automatic categorization, urgency prediction, and email response creation by combining a Decision Tree and the Gemma large language model. However, these frameworks only work for single-label classification and urgency prediction as separate tasks.

45 Other researchers have investigated more sophisticated deep learning algorithms for multi-label text classification. Wang et al. (2024) came up with the BEAL algorithm, which showed that active learning can significantly reduce labeling while retaining good classification results. Similarly, Jain et al. (2024) introduced the Transformer network with knowledge distillation for multi-label document classification and reached very high predictive accuracy on large-scale financial datasets. Nevertheless, both papers ignored the problem of email priority prediction and did not cover multi-task prediction within one architecture.

Some researchers concentrated on improving the email workflow by means of a rule-based approach and RPA. Singh et al. (2023) created an online rule-learning tool called EmFore for automatic classification of email folders, while Khare et al. (2022) suggested an RPA-based assistant that will perform some automated operations with emails. Even though both approaches provide increased efficiency, they depend on rule-based or process automation and do not use contextual representation of language for predicting labels and priority at once.

15 Additionally, comprehensive literature reviews have provided more insight into current research gaps. The literature review by Jáñez-Martino et al. (2022), Kaddoura et al. (2022), Ahmed et al. (2022), and Salloum et al. (2022) shows that most of the existing research still focuses on spam or phishing detection based on binary classification approaches. As such, there have been relatively few efforts made towards developing an intelligent email management system that could classify multiple semantic classes of emails while automatically assigning priorities.

Thus, based on the available literature, three important research gaps can be defined. First, most of the existing studies apply single-label email classification rather than multi-label email classification. Secondly, the problem of priority prediction is usually viewed independently from the classification task. Lastly, there have been relatively few attempts to use explainable artificial intelligence within the prediction process. These research gaps have led to the proposed DistilBERT multi-task model presented in this paper.

3. Methodology

In this research, we introduce a framework for automating email management through multi-label email classification and prioritization in an integrated manner via a shared DistilBERT encoder. The process is depicted in Figure 1 below.

Figure 1. Overall Framework of the Proposed System

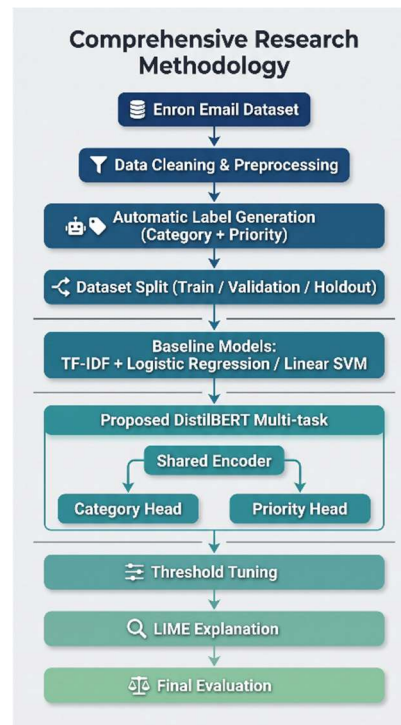


Figure 1 shows the complete pipeline of the proposed approach. This starts with data pre-processing for the Enron Email Dataset and label creation for categories and priorities. The dataset is then split into train, validation, and hold-out sets. The baseline ML approaches as well as the proposed DistilBERT multi-task approach are then built and evaluated. Lastly, threshold tuning and LIME are utilized to enhance classification results and make predictions interpretable.

3.1 Dataset and Data Preparation

The experiments have been performed with the use of the Enron Email Dataset, which is one of the most popular benchmark datasets for email classification studies. Given that there are no categories or priorities labeled

in the initial dataset, a special technique for automatic labeling of categories and priorities of email messages has been implemented.

Preprocessing of email messages included such stages as deletion of duplicate messages, merging of email subject and body, deletion of unnecessary characters and symbols, and elimination of too short messages. As a result, there have been created 190,280 email messages, ready for further modeling.

In order to evaluate models' performance fairly, the clean dataset has been split into training, validation, and testing parts (70/10/20%) with stratified sampling.

Table 1. Dataset Partitioning

Dataset	Samples
Training	35,000
Validation	5,000
Holdout Test	10,000

Table 1 shows the dataset used for model training and evaluation. The validation data was used to optimize the model's parameters and thresholds, and the holdout test data were kept out of the model training process entirely for an unbiased evaluation of model generalization.

3.2 Baseline Models

To provide a good reference point against which the proposed framework could be compared, two standard machine learning models have been used as a basis to develop the proposed method. These models utilized the Term Frequency–Inverse Document Frequency (TF-IDF) approach in converting emails into numerical feature vectors.

For multi-label email categorization, Logistic Regression was used in conjunction with a one-vs-rest approach, wherein an individual binary classifier was developed for each label. This is considered a good baseline model for multi-label text classification owing to its simplicity and efficient computations, not to mention that it is a common tool used in Natural Language Processing (NLP).

For priority prediction, the Linear SVM classifier was utilized as the baseline model. Linear SVM has shown excellent performance in high-dimensional

text classification tasks and will be used as the baseline against which the deep learning model can be compared.

The models were trained on the same training data and evaluated against the same holdout test set that was used for the proposed method.

Figure 2. Baseline Machine Learning Pipeline

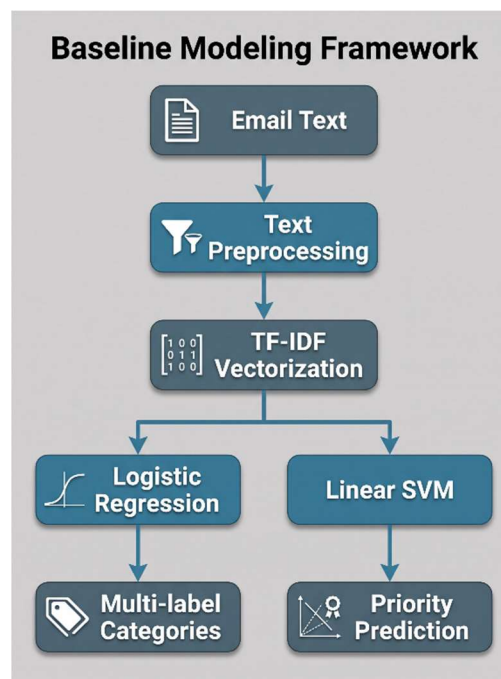


Figure 2 shows the flow of work for the baselines. Post pre-processing, each email is represented as a feature vector using the TF-IDF method. Logistic Regression is utilized for multi-label category classification, whereas Linear SVM is used for predicting priority. The performance of these models serves as the benchmark for evaluating the proposed DistilBERT model.

3.3 Proposed DistilBERT Multi-task Model

This framework is built on DistilBERT, a lighter version of the transformer-based model that keeps the language understanding capacity of BERT and demands less computing power. In contrast to baseline models, this model learns contextual features from the raw data of the emails instead of the hand-crafted TF-IDF features.

To predict email categories and priorities simultaneously, a multitask learning model design is used where the input email is first encoded with the help of a

shared encoder of DistilBERT, and then the learned context is fed into two different predictors.

The first predictor performs multi-label classification for email category prediction with the use of Sigmoid activation, thereby allowing an email to be associated with multiple categories at the same time. The second predictor performs priority prediction with the help of the SoftMax classifier, where each email gets mapped to one of the predetermined priority classes.

Figure 3. Proposed DistilBERT Multi-task Architecture

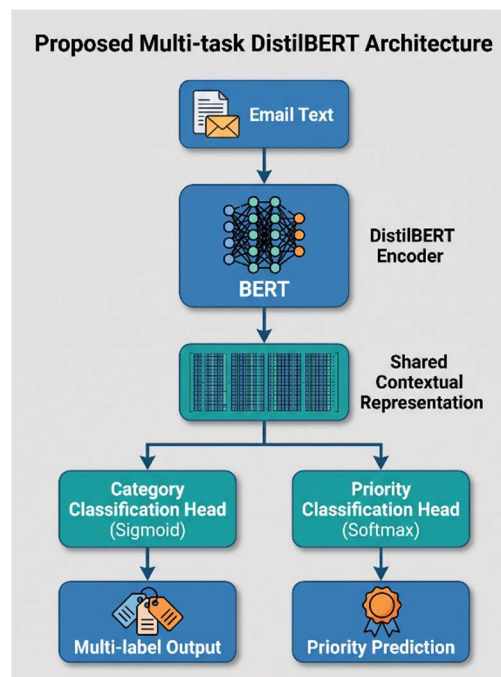


Figure 3 shows the architecture of the suggested multi-task learning model. The DistilBERT encoder shares semantic information obtained from the input emails, which is used by two different output heads for classification and priority prediction of the tasks simultaneously. With this single architecture, there is no need for separate models, and both tasks can benefit from a common linguistic representation.

3.4 Experimental Setup

All experimental work was performed with the help of the Python language and the PyTorch library for deep learning models. The developed DistilBERT model was fine-tuned on the preprocessed dataset of emails using the AdamW

optimizer and mixed precision training. The training process took place on the training subset while the validation subset was used for performance analysis, hyperparameter optimization, and avoiding overfitting. The independent test hold-out subset stayed untouched during the training and optimization phases and was used only for performance evaluation.

Table 2. Training Configuration of the Proposed DistilBERT Model

Parameter	Value
Optimizer	AdamW
Learning Rate	2×10^{-5}
Batch Size	16
Number of Epochs	5
Mixed Precision Training	Enabled
Encoder	DistilBERT

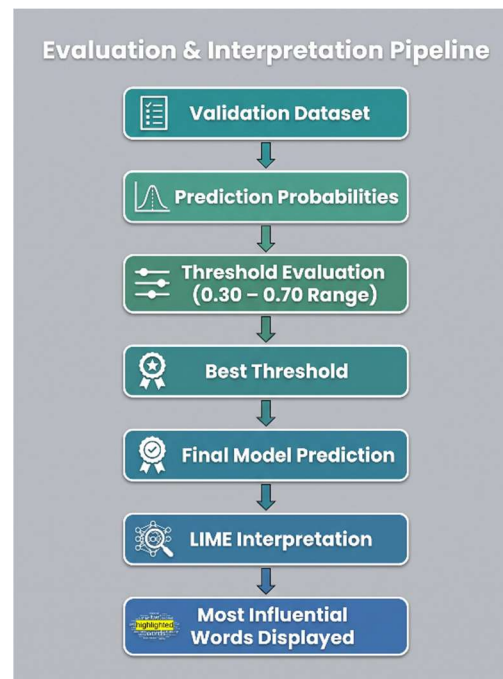
Hyperparameters used for fine-tuning the DistilBERT model are highlighted in Table 2. They were chosen in order to guarantee convergence and efficient training of the model.

3.5 Threshold Tuning and Explainable AI

After training the model, threshold tuning was done to refine multi-label category classification. Instead of applying a conventional classification threshold of 0.50, other thresholds were considered on the basis of the performance of the model on the validation set. The threshold providing the best trade-off between Macro F1-Score and Hamming Loss was chosen and then used for the evaluation on the holdout test set.

For better interpretation of the results provided by the proposed approach, the Local Interpretable Model-Agnostic Explanations (LIME) method was incorporated into the model after its training. LIME analyzes the contribution of each word to the final output of the model and thus shows how particular words affect the decisions of the model.

Figure 4. Post-training optimization and explanation workflow.



This is depicted in Figure 4. The prediction probability of the trained DistilBERT model was tested using different threshold levels on the validation dataset to determine the best threshold level for performing multi-label classification. The interpretation of the final prediction was carried out through LIME to indicate which words had the most contribution to the model's decision.

3.6 Evaluation Metrics

Performance analysis of the proposed model as well as the baseline models was done using performance measures that have been widely used in multi-label text classification and multi-class classification. For the category prediction problem, measures such as Subset Accuracy, Micro F1-score, Macro F1-score, Hamming Loss, and Jaccard Score were used. On the other hand, Accuracy, Precision, Recall, and Macro F1-score measures were utilized for the prediction of priority classes with special focus on Recall for the High Priority category, as it is very important in classifying urgent emails.

Use of the same evaluation criteria for all models with the same hold-out test set ensured an unbiased comparison of the proposed DistilBERT model with the baseline machine learning approaches.

4. Results and Discussion

4.1 Performance Comparison of Category Classification

The DistilBERT multi-task model that was suggested was compared against the baseline Logistic Regression model for the multi-label classification problem of email categories by utilizing the same holdout test set and metrics of assessment. Specifically, the following were used: Macro F1-Score, Micro F1-Score, Hamming Loss, Subset Accuracy, and Jaccard Score. These results are presented in Table 3.

Table 3. Category Classification Performance Comparison

Metric	Logistic Regression	DistilBERT
Macro F1-score	0.8751	0.9672
Micro F1-score	0.8711	0.9681
Hamming Loss	0.0691	0.0175
Subset Accuracy	0.6840	0.9303
Jaccard Score	0.8280	0.9366

The performance of the baseline Logistic Regression model and the proposed DistilBERT model in categorizing emails is presented in Table 3. The proposed method proved to be better than the baseline in all measures, thereby showing better abilities in multi-label classification of emails.

Figure 5. Comparison of Macro F1-score for category classification.

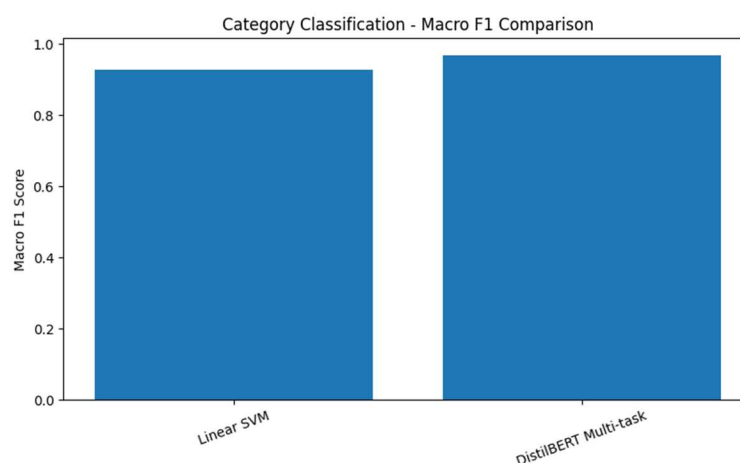


Figure 5 shows the Macro F1-score produced by the baseline Logistic Regression model and the proposed DistilBERT model. The high F1-score in

the case of DistilBERT is due to its better capacity to capture the semantic context information, thus achieving higher accuracy in multi-label classifications of emails.

The Macro F1-score of the proposed model is 0.9672, which is a huge improvement compared to the baseline Logistic Regression model whose Macro F1-score is 0.8751. Improvements are also noted in other metrics of evaluation.

Hamming Loss decreases from 0.0691 to 0.0175, showing a significant decrease in individual label prediction errors. At the same time, there is a big increase in the Subset Accuracy of the proposed model, showing an increased number of completely correct multi-label predictions.

The reason for such improvements in performance lies in the contextual language representations learned by DistilBERT. In contrast to the TF-IDF approach, which mostly focuses on the frequencies of the words, DistilBERT learns semantic relations between the words and understands their context. As a result, it is easier to detect multiple topics in a single email.

4.2 Performance Comparison of Priority Prediction

The effectiveness of the proposed DistilBERT multi-task model in solving the email priority prediction problem was also measured and compared to that of the Linear Support Vector Machine (Linear SVM) – the baseline model. Both models have been trained and tested on the same partitions of the dataset.

In Table 4, the performance of both models is shown using Accuracy, Macro F1-score, and Recall for the High Priority class.

Table 4. Priority Prediction Performance Comparison

Metric	Linear SVM	DistilBERT
Accuracy	0.9079	0.9529
Macro F1-score	0.8695	0.9327
High Priority Recall	0.6900	0.8627

Table 4 provides a performance comparison between the baseline Linear SVM model and the proposed DistilBERT multi-task model in terms of priority prediction. The DistilBERT model outperforms the Linear SVM model in all performance measures.

Figure 6. Priority Prediction Performance

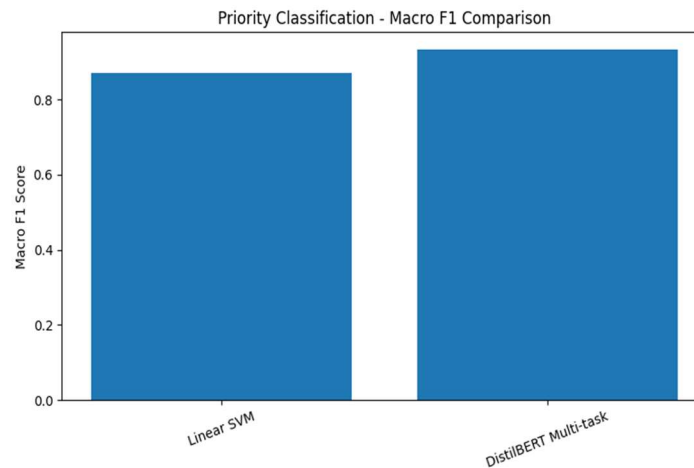


Figure 6 illustrates the Macro F1-score attained by the Linear SVM baseline and the proposed DistilBERT model for predicting priorities. This better score attained by the DistilBERT demonstrates a higher classification performance for all the priority classes, both majority and minority classes..

Figure 7. High Priority Recall Comparison

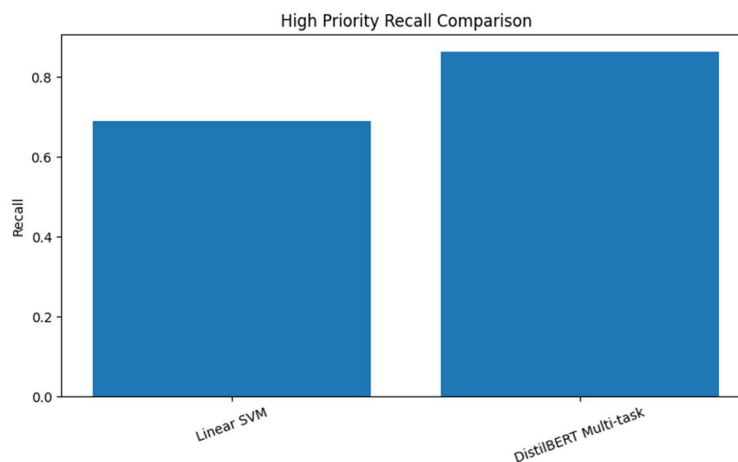


Figure 7 demonstrates the recall obtained for the High Priority class. The suggested DistilBERT architecture showed higher results of recall than the Linear SVM baseline, which means its capability of determining urgent emails increased significantly.

As was already mentioned, the DistilBERT architecture gave 95.29% accuracy for the task, while the Linear SVM baseline produced only 90.79%. The increase in the Macro F1 score was observed as well – 0.8695 vs 0.9327.

One of the most significant improvements is the growth of the recall of the High Priority class – from 0.6900 to 0.8627. It is one of the most important results since it shows that the failure to recognize the urgency of the message can cause delays in the processing of critical and urgent requests. Higher recall of DistilBERT means that it recognizes urgent emails significantly better and is capable of good classification as a whole.

Such an increase in the results was caused by the contextual language understanding implemented in DistilBERT, which enables it to distinguish urgency- and priority-related linguistic peculiarities of the email content without statistical representations of words.

4.3 Overall Model Comparison

In order to present an evaluation of the entire framework presented in this work, comparisons were made with regard to the performances of both the baseline models and the multi-task model using DistilBERT in both classification tasks. Table 5 presents a summary of the entire experiment results.

Table 5. Overall Comparison Between Baseline Models and the Proposed Model

Model	Category Macro F1	Hamming Loss	Priority Accuracy	Priority Macro F1	High Recall
Linear Models (Logistic Regression + Linear SVM)	0.9269*	0.0370*	0.9079	0.8695	0.6900
Proposed DistilBERT Multi-task	0.9672	0.0175	0.9529	0.9327	0.8627

* The category values are the same as those from the final Linear SVM category classification results presented in this thesis.

Table 5 shows a summary of the performance comparisons between the baseline machine learning models and the proposed DistilBERT multi-task model.

4.4 Discussion

Experimental results showed that the proposed framework of using DistilBERT for multi-tasking outperformed the baseline machine learning models when applied to both email category classification and priority prediction tasks. This result can be explained by the use of contextualized representations of languages that can recognize not only word frequency statistics but also relationships between words.

In the case of multi-label category classification, the model outperformed traditional TF-IDF models in terms of achieving a higher value of Macro F1-score and a lower value of Hamming Loss. This result implies that the model was better at identifying several email categories at once and avoiding incorrect labels. Such a result is especially valuable because most emails discuss several administrative issues at once.

Likewise, the model performed better in the task of predicting email priorities. In particular, the increase in the recall value of the High Priority class means that the model was more efficient at detecting high priority emails. In a practical sense, such an achievement implies that fewer high priority emails will be misinterpreted and left unanswered.

Another key contribution of this research is the use of a multi-task learning model architecture. Different from the baseline technique that uses separate models for the category classification and the priority prediction tasks, the proposed approach carries out both tasks through a common DistilBERT encoder. This design not only makes the whole model simpler but also lets both tasks utilize shared contextual embeddings extracted from the emails.

Another interesting contribution of this research is the improvement of the multi-label classification performance through post-training threshold tuning without changing the model architecture or retraining. With the optimal classification threshold selected from the validation data, the proposed framework provides better classification performance.

In the end, the inclusion of the LIME explainability method led to greater transparency of the system due to the determination of the most impactful words. It was confirmed through explanations that the model used relevant

words and phrases associated with the content of the emails instead of unrelated words, ensuring confidence in the decisions made by the model.

In general, these results show that the use of contextual language models, multi-task learning, optimal threshold setting, and explainable artificial intelligence allows developing an efficient intelligent system for email management with the ability to automatically classify and predict the priority of the emails.

5. Conclusion

The paper has presented a unified approach to managing emails through a DistilBERT multi-task learning framework that is able to perform both multi-label email categorization and priority prediction at the same time.

The effectiveness of the suggested approach has been experimentally verified by applying it to the Enron Email Dataset and comparing its results with those obtained through the application of the traditional machine learning algorithms, which utilized TF-IDF features.

Experimental results have shown that the use of DistilBERT resulted in an improvement in classification performance for both classification tasks. In particular, concerning the category classification task, DistilBERT achieved better Macro F1-score values and significantly reduced Hamming Loss compared to the baseline algorithms.

Concerning the task of priority prediction, DistilBERT demonstrated an increase in overall accuracy, Macro F1-score, and, more importantly, the recall of the High Priority class, which is of high practical significance for email management.

Apart from performance improvement, the experiment has shown that threshold tuning may be used to optimize multi-label classification. In addition, integration of LIME into the system provides interpretability of individual predictions.

Future Work

Future research could explore the utilization of large-scale multilingual transformer networks for multilingual email classification. Another way to

enhance the performance of the classification system could be through the integration of attachment analysis, sender information, and conversation context. Evaluation of the proposed approach on other real-life datasets could provide information about its generalizability.

48

References

Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi, B., & Shah, T. (2022). Machine learning techniques for spam detection in email and IoT platforms: Analysis and research challenges. *Security and Communication Networks*, 2022, Article 1862333. <https://doi.org/10.1155/2022/1862333>

Bhalla, S., Kaur, M., & Sharma, V. (2021). Automatic genre categorization of emails into predefined categories. *Expert Systems with Applications*.

Chuttur, M. Y., & Parianen, Y. (2022). A comparison of machine learning models to prioritise emails using emotion analysis for customer service excellence. *Knowledge Engineering and Data Science*, 5(1), Article 4. <https://doi.org/10.17977/um018v5i12022p41-52>

Duong, T., Tran, A., Dang, V., Tran, T., Le, H.-S., & Nguyen, P. (2025). AEEM: A deep learning approach to automated educational email management. *Journal of Information Hiding and Multimedia Signal Processing*, 16(2).

Fareed, B., Din, G. M., Khan, M. R. A., Fatima, T., & Shahid, S. (2025). Machine learning-based email spam detection: Accuracy, overfitting and robustness analysis. *European Journal of Applied Science, Engineering and Technology*, 3(6), 99–108. [https://doi.org/10.59324/ejaset.2025.3\(6\).06](https://doi.org/10.59324/ejaset.2025.3(6).06)

Hirpara, D., & Polara, V. (2025). InboxG: Smart mail classifier. *International Journal on Science and Technology (IJSAT)*, 16(2).

Iqbal, M., & Khan, A. (2022). Email classification analysis using machine learning techniques. *Data Science Journal*.

Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., & Alegre, E. (2022). A review of spam email detection: Analysis of spammer strategies. *Progress in Artificial Intelligence*.

Kaddoura, S., Chandrasekaran, G., Popescu, D. E., & Duraisamy, J. H. (2022). A systematic literature review on spam content detection and classification. *PeerJ Computer Science*, 8, e830. <https://doi.org/10.7717/peerj-cs.830>

Khare, N., Singh, S., Mishra, R., Prakash, S., & Dixit, P. (2022). E-mail assistant: Automation using robotic process automation (RPA). In *2022 International Conference on Decision Aid Sciences and Applications (DASA)* (pp. 511–516). IEEE.

Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2022). A systematic literature review on phishing email detection using natural language processing techniques. *Applied Sciences*.

Singh, M., Cambronoero, J., Gulwani, S., Le, V., & Verbruggen, G. (2023). EmFore: Online learning of email folder classification rules. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)* (pp. 4262–4273). ACM. <https://doi.org/10.1145/3583780.3614863>

Wang, L., Zhang, H., Zhang, W., Dai, L., Liang, Y., & Shi, H. (2024). Deep active learning for multi-label text classification. *Neural Computing and Applications*.