

ابوبكر حسين

%D8%A7%D9%84%D9%88%D8%B1%D9%82%D8%A9
%D8%A7%D9%84%D8%A8%D8%AD%D8%AB%D9%8A%D8%A9...

 Midocean University

Document Details

Submission ID

trn:oid:::2945:404408533

Submission Date

Aug 1, 2026, 1:35 PM GMT+3

Download Date

Aug 1, 2026, 5:13 PM GMT+3

File Name

الورقة البحثية بعد التعديل الاول.xcod.

File Size

191.3 KB

30 Pages

7,145 Words

43,297 Characters

17% Overall Similarity

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

Match Groups

-  **107 Not Cited or Quoted** 16%
Matches with neither in-text citation nor quotation marks
-  **5 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

- 10%  Internet sources
- 9%  Publications
- 13%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

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

- 107 Not Cited or Quoted 16%**
Matches with neither in-text citation nor quotation marks
- 5 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

- 10% Internet sources
- 9% 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** Internet
www.teses.usp.br <1%
- 2** Publication
Victor Jima. "Integrating Traditional SMEs into Hybrid Digitalized Trade Navigatio... <1%
- 3** Internet
www.jihmsp.org <1%
- 4** Internet
assets.pubpub.org <1%
- 5** Submitted works
University of Northumbria at Newcastle on 2024-08-28 <1%
- 6** Internet
ojs.bonviewpress.com <1%
- 7** Publication
"AI Technologies for Information Systems and Management Science", Springer Sc... <1%
- 8** Internet
arxiv.org <1%
- 9** Submitted works
Middle East College on 2025-11-14 <1%
- 10** Internet
www.ijraset.com <1%

11	Submitted works	British Academy of Jewellery on 2025-07-17	<1%
12	Submitted works	University of Warwick on 2023-09-05	<1%
13	Internet	www.turcomat.org	<1%
14	Submitted works	Al Hussein bin Talal University on 2024-01-08	<1%
15	Submitted works	Strathmore University (Main Account) on 2026-03-25	<1%
16	Internet	ris.utwente.nl	<1%
17	Publication	Muzammil, Nazeefa. "Sensi-Train: Context-Aware Sensitive Content Detection Usi...	<1%
18	Publication	Bhabani Prasanna Pattanaik, Dharendra Kumar Shukla, Sambit Satpathy, Kamala...	<1%
19	Publication	Harvinder Singh, Priyanka Kaushal, Sarabjeet Kaur. "Advanced Computing and AI...	<1%
20	Submitted works	Liverpool John Moores University on 2023-05-29	<1%
21	Internet	ir.nust.na	<1%
22	Internet	pure.nwpu.edu.cn	<1%
23	Internet	www.kaggie.com	<1%
24	Submitted works	Southern Cross University on 2024-05-12	<1%

25	Publication	Bishan Yang, Jian-Tao Sun, Tengjiao Wang, Zheng Chen. "Effective multi-label acti...	<1%
26	Submitted works	Liverpool Hope on 2023-01-22	<1%
27	Submitted works	University of Hull on 2026-05-07	<1%
28	Submitted works	Curtin University LTI on 2026-05-19	<1%
29	Publication	Lecture Notes in Computer Science, 2016.	<1%
30	Submitted works	TAFE NSW Higher Education on 2024-09-08	<1%
31	Publication	Poonam Nandal, Tapas Kumar, Meeta Singh, Mamta Dahiya. "Progressive Compu...	<1%
32	Submitted works	University College London on 2021-05-03	<1%
33	Submitted works	University of Hertfordshire on 2026-06-23	<1%
34	Submitted works	York St John University on 2025-09-18	<1%
35	Internet	arcabc.ca	<1%
36	Submitted works	Queen Mary and Westfield College on 2025-08-23	<1%
37	Submitted works	West Virginia University on 2024-10-09	<1%
38	Internet	repo.pw.edu.pl	<1%

39	Submitted works	Southern Cross University on 2024-05-28	<1%
40	Submitted works	University of Birmingham on 2024-08-28	<1%
41	Submitted works	University of Westminster on 2026-04-17	<1%
42	Internet	link.springer.com	<1%
43	Internet	www.irrodl.org	<1%
44	Publication	"Computational Science and Its Applications – ICCSA 2026 Workshops", Springer S...	<1%
45	Submitted works	Islington College,Nepal on 2026-07-31	<1%
46	Submitted works	KJSCE on 2026-04-13	<1%
47	Internet	aclanthology.org	<1%
48	Publication	özdemir, Mehmet Akif. "Image-Based Representations of Physiological Signals: D...	<1%
49	Submitted works	Dublin Business School on 2025-08-27	<1%
50	Publication	Ephrem Admasu Yekun, Abrahaley Teklay Haile. "Student Performance Prediction...	<1%
51	Publication	Kailai Yang, Tianlin Zhang, Shaoxiong Ji, Sophia Ananiadou. "A Bipartite Graph is ...	<1%
52	Submitted works	Liverpool John Moores University on 2021-01-16	<1%

53	Submitted works	University of East London on 2026-05-08	<1%
54	Submitted works	University of Essex on 2026-07-23	<1%
55	Internet	medicaljournalshouse.com	<1%
56	Internet	www2.mdpi.com	<1%
57	Publication	Abhay K. Srivastava, Lovish Bhatia, VS Pandey, Ajay K. Sharma. "Forensically-integ...	<1%
58	Submitted works	Berlin School of Business and Innovation on 2025-01-31	<1%
59	Submitted works	Higher Education Commission Pakistan on 2026-06-22	<1%
60	Submitted works	Innopolis University on 2020-05-12	<1%
61	Submitted works	JB Institute of Engineering And Technology on 2026-04-08	<1%
62	Publication	Lecture Notes in Computer Science, 2011.	<1%
63	Publication	Lecture Notes in Computer Science, 2014.	<1%
64	Publication	Lecture Notes in Computer Science, 2015.	<1%
65	Submitted works	Mutah University in Jordan on 2021-12-27	<1%
66	Publication	Rishabha Malviya, Shivam Rajput. "Artificial Intelligence for Neural Health - Diagn...	<1%

67	Submitted works	Tilburg University on 2026-06-29	<1%
68	Submitted works	University of East London on 2025-09-08	<1%
69	Submitted works	University of Hertfordshire on 2026-04-26	<1%
70	Submitted works	University of Northumbria at Newcastle on 2023-09-12	<1%
71	Submitted works	University of Sheffield on 2019-08-27	<1%
72	Internet	direct.mit.edu	<1%
73	Internet	papers.academic-conferences.org	<1%
74	Internet	researchid.co	<1%
75	Internet	www.kolabtree.com	<1%
76	Internet	www.uyik.org	<1%
77	Publication	Abdelkader Berrouachedi, Rakia Jaziri, Gilles Bernard. "A deep ensemble approac...	<1%
78	Submitted works	Al Balqa Applied University on 2025-11-28	<1%
79	Publication	Bland, Ashley. "Optimizing Machine Learning for Real-Time FNIRS-Based Prosthet...	<1%
80	Submitted works	Dublin City University on 2025-07-31	<1%

81	Submitted works	
Liverpool John Moores University on 2021-03-12		<1%
82	Submitted works	
Universiti Tun Hussein Onn Malaysia on 2026-05-25		<1%
83	Submitted works	
University of Central Lancashire on 2023-01-04		<1%
84	Submitted works	
University of Texas Health Science Center on 2025-12-09		<1%
85	Publication	
"Smart Data Intelligence", Springer Science and Business Media LLC, 2024		<1%

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

Abstract

Email management is getting more difficult as organizations receive increasing amounts of complex communications. In this paper, we propose a multi-task learning approach with a DistilBERT encoder for concurrent multi-label email classification and priority prediction. The model was trained and evaluated on a stratified sample of 50,000 emails selected from the Enron Email Dataset. Since the original Enron dataset does not contain predefined labels, we developed an automated labeling strategy using domain-specific keyword dictionaries to generate 6 functional category labels (Administration, Financial, HR, Legal, Market, and Other) and 3 priority classes (High, Normal, and Low). The quality of these labels was validated through label distribution analysis and category co-occurrence metrics. 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 for priority prediction. Moreover, thresholding at 0.60 and LIME-based interpretability further enhanced the system's reliability.

المخلص

تزداد صعوبة إدارة البريد الإلكتروني مع زيادة حجم الاتصالات التنظيمية وتعقيدها. تقترح هذه الدراسة نموذج تعلم متعدد المهام يعتمد على مشفر DistilBERT للتصنيف متعدد الوسوم والتنبؤ بالأولوية في آن واحد. تم تدريب النموذج على عينة طبقية مكونة من 50,000 رسالة من مجموعة بيانات Enron. وبما أن البيانات الأصلية غير موسمة، فقد تم توليد الوسوم آلياً لـ 6 فئات وظيفية و3 مستويات للأولوية. أظهرت النتائج تحقيق دقة 95.29% للتنبؤ بالأولوية ودرجة Macro F1 بلغت 0.9672 لتصنيف الفئات، مع تعزيز الشفافية باستخدام الذكاء الاصطناعي القابل للتفسير (LIME).

Keywords: Multi-task Learning, DistilBERT, Multi-label Email Classification, Priority Prediction, Explainable AI, LIME, Natural Language Processing

1. Introduction

Organizing emails according to their level of relevance and urgency has become a time-consuming process, resulting in procrastination (Khare et al., 2022). The current research highlights that the current approach of email categorization is not feasible for contemporary organizations (Iqbal & Khan, 2022; Bhalla et al., 2021). In addition, the use of deep learning for organizing emails in education and business has proven to be beneficial (Duong et al., 2025; Hirpara & Polara, 2025).

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).

To fill in the gaps mentioned, the study attempts to answer the following research questions:

RQ1: What is the effectiveness of using a multitasking DistilBERT algorithm to classify emails according to their functionalities?

RQ2: What is the accuracy of predicting email priority levels by the same architecture?

RQ3: How can explainable AI (LIME) be used to make the text-based decision transparent?

Based on those answers, the main contributions of this study will include:

1. Unified Multi-task Learning Framework using a shared DistilBERT encoder for efficient training
2. Automation of Labelling to deal with unlabeled emails from companies
3. Threshold Optimization for Multi-label Classification
4. Explanation of Model Predictions using LIME.

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.

The selection of Logistic Regression for multi-label category classification is because of the stability of the model in high-dimensional text space with the one-vs-rest technique, and Linear SVM for priority prediction due to the efficiency of the model in multi-class text classification.

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 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.

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.

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.

Table 1: Comparative Analysis of Literature Review Studies

Author (Year)	Paper title	Description	Dataset	Methodology / Technique	Results	Advantages	Disadvantages
Ahmed et al. (2022)	ML Techniques for Spam Detection in Email and IoT Platforms	Challenges of spam in Email and Internet of Things (IoT).	Analytical Review	ML Algorithms Comparison	Identified research gaps in securing IoT devices from spam.	Links Email security with the modern IoT ecosystem.	Very broad scope; lacks depth in Email-specific solutions.
Bhalla et al. (2021)	Automatic Genre Categorization of Emails into predefined categories	Classifying emails by genre (Official, Personal, Business).	Enron and others	TF-IDF + SVM	Accuracy exceeds 90%.	Simplicity and high execution speed.	Rely heavily on predefined keyword lists.
Chuttur et al. (2022)	A Comparison of Machine Learning Models to Prioritize Emails using Emotion Analysis	Prioritizing emails based on customer emotion analysis.	Customer Service datasets	Emotion Analysis + (SVM, Naive Bayes, Random Forest)	Random Forest achieved the best accuracy at 84%.	Integrates emotional context into priority queuing.	Limited scope (focused only on customer service).
Duong et al. (2025)	AEEM: A Deep Learning Approach to Automated Educational Email Management	System for managing academic emails for students.	University-specific data (Vietnam)	Deep Learning + Text Summarization	93% accuracy with significant improvement in class balance.	Combines summarization with classification for better UI.	Localized dataset; might not generalize to other languages.
Fareed et al. (2025)	ML-Based Email Spam Detection: Accuracy, Overfitting, and Robustness	Analyzing model robustness and resistance to overfitting.	Public Ham/Spam Dataset	19 Classifiers + Robustness Analysis	Random Forest was found to be the most stable and robust.	Extensive comparison across 19 different algorithms.	Focuses mostly on classic ML rather than Neuro-AI.
Hirpara & Polara (2025)	InboxG – Smart Mail Classifier	An AI system for manufacturing companies to detect urgency.	Elecon Engineering Co. Data	Gemma 1.1 2B (LLM) + Decision Tree	Accurate urgency detection and intelligent autoreplies.	Utilizes modern Large Language Models (LLMs).	Requires high hardware specs for LLM inference.
Iqbal & Khan (2022)	Email classification analysis using machine learning techniques	Feature selection analysis to improve spam detection.	UCI Spambase	Point-Biserial Correlation + SVM & ANN	98.06% accuracy using Artificial Neural Networks (ANN).	Feature selection methodology.	Uses a relatively old dataset (UCI Spambase).

26	39	Jáñez-Martino et al. (2022)	A review of spam email detection: analysis of spammer strategies	Comprehensive review of spammer strategies and the "Dataset Shift" problem.	SLR (Systematic Literature Review)	SLR Methodology	Identified "Dataset Shift" as the main challenge for filter robustness.	In-depth analysis of adversarial behavior (spammers).	Theoretical study lacks new experimental benchmarking.
20		Kaddoura et al. (2022)	A systematic literature review on spam content detection and classification	Review of spam in social media (FB, Twitter, YT) and Email.	Multi-platform review	SLR	Email remains the most targeted medium for spam content.	Compares spam behavior across different platforms.	Broad focus reduces specific technical depth for email.
		Khare et al. (2022)	E-Mail Assistant – Automation using Robotic Process Automation (RPA)	Automating email handling (attachments, sorting, replies).	Experimental setup	RPA (UiPath) + AI	Successfully reduced human effort in routine email tasks.	Practical integration with corporate office environments.	Dependent on commercial tools (UiPath).
12	12	Salloum et al. (2022)	A Systematic Literature Review on Phishing Email Detection Using NLP	Systematic review of NLP techniques used in detecting phishing emails.	SLR (2006-2021 period)	SLR + NLP Techniques Analysis	DL models are starting to outperform traditional ML in URL/link detection.	High comprehensiveness across a long timeframe.	Does not cover the latest Transformer-based models (2023+).
40		Singh et al. (2023)	EmFore: Online Learning of Email Folder Classification Rules	Continuously learning folder classification rules.	Microsoft Internal Data	Online Learning + Rule-based	Continuous improvement in folder filing accuracy over time.	Adapts dynamically to changing user behavior.	Requires a constant stream of data to remain effective.
77		Wang et al. (2024)	Deep active learning for multi-label text classification	Deep active learning for multi-label classification.	Reuters, AAPD (Benchmark)	Bayesian DL + Active Learning (BEAL)	Reduced labeled data requirements by 30% with stable accuracy.	Significantly lowers manual labeling costs.	High computational complexity of Bayesian models.

Table 1 provides a comparative analysis of some seminal works in the field of email classification, spam detection, priority prediction, and automated email management. It is important to note that it gives a brief insight into the contributions of previous research in terms of methodology, datasets used, results, and limitations in the field of email classification, which started from classical machine-learning algorithms like TF-IDF, SVM, and ensembles to more recent deep-learning and transformer-based models. This table presents the research gap in this paper since no

unified multi-task transformer-based system for multi-label email classification and priority prediction has been developed.

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.

The proposed architecture attempts to bridge the multi-tasking gap by using a common DistilBERT encoder that enables learning of common semantic features used in classification and prioritization, while LIME solves the black-box problem of transformers.

3. Methodology

This study employs a reproducible approach with a clear methodology that goes through several clearly defined stages from data collection to the ultimate evaluation of the proposed model. The dataset used as the data source for this research is the Enron Email Dataset, which provides numerous real-life emails. The preprocessing step is carried out to make sure that the quality of the data is adequate, including duplicate removal, resolving inconsistent encodings, eliminating automatic signatures, and textual field normalization.

Then, automatic label generation takes place in order to generate category and priority labels according to linguistic rules and context criteria. After that, the data is divided into training, validation, and testing sets, with a stratified split of 70%, 15%, and 15% correspondingly.

Baseline performance is determined with the help of traditional machine learning models that include TF-IDF in conjunction with logistic regression and Linear SVM. In turn, the suggested approach involves the use of a multi-task architecture that includes DistilBERT and is comprised of an encoder part and two classifiers for categories and priorities, respectively. Training is done by means of AdamW optimization with a learning rate equal to 2×10^{-5} , a batch size of 32, and five epochs.

Thresholding is carried out on the validation set with the aim of achieving the optimum combination of precision and recall. Interpretability is provided via LIME and is meant to identify the most significant features of the text involved in the process of prediction. Evaluation of final results is done by means of precision, recall, and F1-score.

3.1 Dataset and Data Preparation

The experiments were conducted using the Enron Email Dataset, a widely used benchmark corpus for email classification research. Since the dataset does not include predefined category or priority annotations, a structured automatic labeling framework was developed to generate the required labels. The category taxonomy consisted of six classes—Administration, Financial, HR, Legal, Market, and Other—defined based on common organizational communication themes. Priority levels were organized into three classes: High, Normal, and Low.

Label generation was performed using weighted keyword dictionaries. Each category was associated with a curated list of domain-specific keywords (e.g., *budget*, *invoice*, and *audit* for the Financial category). For every email, category scores were computed by summing the weights of matched keywords, allowing multiple categories to be assigned when several domains were strongly represented. Priority labels were generated using a separate dictionary containing urgency-related expressions (e.g., *urgent*, *ASAP*, *deadline*), where the highest accumulated score determined a single priority class.

To ensure reproducibility, explicit rules governed the labeling process, including minimum score thresholds, normalization of repeated keyword occurrences, and exclusion of boilerplate text such as signatures and disclaimers. Examples of labeled emails were reviewed to verify that the assigned labels reflected the semantic content of the messages.

In total, the preprocessing pipeline generated 190,280 clean emails by imposing several exclusion criteria. In particular, duplicate emails, null emails, and emails having fewer than five words were excluded from the dataset. Moreover, emails having more than 512 tokens were also excluded to

enable the utilization of transformer-based architecture and also avoid very long outlier emails, which can create unnecessary noise in the modeling process. Emails were stripped of their subject and body fields; non-informative characters and boilerplate text such as sign-off and disclaimer were also excluded.

While all of the entire clean dataset was available, a stratified sample of 50,000 emails was chosen to be used in the experiment. This sampling methodology was chosen because the purpose was to keep computational costs low during the training process but still retain proportionality in category and priority labeling of emails in the full dataset.

The rest of the cleaned emails were not employed for training since there was a high cost involved in fine-tuning transformer models on such big datasets. The preliminary results have shown that no significant gains in performance were achieved as the size of the dataset increased above 50,000 emails. For this reason, the dataset of 50,000 emails was chosen for the experiment as it was balanced and representative enough.

15 The cleaned dataset was then split into training, validation, and test subsets with a (70/10/20)% allocation, respectively. Due to the nature of the problem, the task of assigning multiple labels per email, standard stratification does not suffice. Iterative stratification (Sechidis et al., 2011), therefore, was performed on the multi-label category-priority vectors.

70 A random seed of 42 was used for reproducibility of the splits. Before any splitting, duplicate and quasi-duplicate emails were filtered out to avoid data leakage. Furthermore, emails from the same conversation thread were identified through normalization and thread identification techniques, and all such emails were assigned to one and only one subset to avoid semantic leakage.

This approach guarantees (1) preservation of the multi-label distributions, (2) consistency of the priority-level ratios in each split, and (3) that none of the emails and threads end up in two or more subsets.

Table 2. Dataset Partitioning

Dataset	Number of Emails	Percentage	Purpose
Training Set	35,000	70%	Model training
Validation Set	5,000	10%	Model selection and threshold tuning
Holdout Test Set	10,000	20%	Final model evaluation
Total Sample	50,000	100%	Experimental dataset

The experimental design used a dataset split that is summarized in Table 2 below. A stratified sampling of 50,000 emails was made out of the whole labeled dataset to improve computational efficiency while maintaining the joint probability distribution of category labels and priority levels. The training data (70%), which was used to train the parameters of the model, the validation data (10%), which was used to tune the hyperparameters and set appropriate thresholds, and the holdout test data (20%), which was not used during any step of the training process, are separated. The class-balance properties of the whole dataset, namely the distribution of categories, priorities, and the number of labels per email, are provided separately.

3.2 Baseline Models

In order to get a reproducible baseline for comparison, the traditional machine learning models have been implemented by TF-IDF vectorization with linear Support Vector Machines (SVMs). For TF-IDF, `ngram_range = (1, 2)`, `max_features = 20,000`, and `min_df = 5` were used to ignore rare words. All the text has been processed by lowercasing and removing English stop-words.

For multi-label category classification, a Linear SVM model has been trained with the One-vs-Rest (OvR) method, in which separate binary classifiers are learned for each category. The baseline model has been implemented by scikit-learn's LinearSVC with a linear kernel, hinge loss, L2 regularization, and `C = 1.0`. As labels are imbalanced, `class_weight = "balanced"` has been used.

For priority prediction, the same TF-IDF vectorizer has been used, and Linear SVM has been trained with scikit-learn's LinearSVC. For the priority model, linear kernel, hinge loss, L2 regularization, and `C = 1.0` have been used. Classes are imbalanced for priority (High = 12.4%, Normal = 51.0%, Low =

36.6%). Therefore, `class_weight = "balanced"` has been used. In addition, LinearSVC does not provide calibrated probability estimates; hence, probability calibration has not been performed.

Both baseline models were trained on the same training split and evaluated on the same holdout test set used for the proposed DistilBERT model.

3.3 Proposed DistilBERT Multi-task Model

The framework is based on the architecture of DistilBERT, an efficient transformer-based model architecture that has a similar language understanding capability to BERT but consumes fewer computational resources. In comparison with baseline architectures, which uses manually crafted TF-IDF features, our model extracts contextual information directly from raw email texts.

The pre-processing of each email involves the use of the DistilBERT tokenizer, in which text is converted to lower case and English stop-words are removed from it. This is followed by WordPiece tokenization and encoding of the email in such a way that its maximum length is equal to 256 tokens, which is specified in the implementation.

Encoded email text is passed to the shared DistilBERT encoder (distilbert-base-uncased), resulting in contextualized hidden states of size 768. These states are pooled into a [CLS,] representation that is passed to two task-specific prediction heads, each of which is preceded by a dropout layer (dropout = 0.1).

The first head conducts multi-label categorization through a linear layer of size $768 \rightarrow 6$ and then uses a Sigmoid activation function that makes it possible for an email to belong to several categories at once.

The second head performs priority prediction with a linear layer of size $768 \rightarrow 3$, then applies a Softmax activation to map an email to the appropriate priority category.

The training process uses a multi-task learning approach in which both heads share the same encoder. Two loss functions are optimized by the model: Binary Cross-Entropy with Logits for category prediction and Cross-Entropy

Loss for priority prediction. The overall loss function is calculated as a weighted average:

$$\mathcal{L}_{total} = 0.5 \cdot \mathcal{L}_{category} + 0.5 \cdot \mathcal{L}_{priority}$$

The proposed architecture makes it unnecessary to develop separate models and lets both tasks profit from shared linguistic representations obtained through DistilBERT.

3.4 Experimental Setup

All experimental runs were performed using Python and the PyTorch library. The DistilBERT model was fine-tuned on the prepared email dataset using AdamW optimization, mixed precision training, and a batch size of 16. The training itself was done on the training subset, while the validation subset was used for tracking performance, tuning of hyperparameters, and prevention of overfitting. The independent hold-out test subset was untouched during the whole training procedure.

The model was trained on an NVIDIA GPU (CUDA-supported environment) using PyTorch 2.x, Transformers 4.x, and Python 3.10. The fixed random seed 42 was set to ensure reproducibility. DistilBERT encoder (distilbert-base-uncased) contains ~66 million parameters, and the entire multi-task model (encoder + two heads) has ~67 million parameters. Training took approximately X hours (replace with your actual time) for 5 epochs with early stopping after 2 consecutive epochs without improvement on validation loss. All results are from a single run due to the fixed seed.

The hyperparameters used during fine-tuning are provided in Table 2. The hyperparameter configuration involves a maximum sequence length of 256 tokens, dropout = 0.1, learning rate = 2×10^{-5} , weight decay = 0.01, optimizer epsilon = 1e-8, and a linear learning-rate scheduler with warmup ratio = 0.1. In order to calculate the multi-task loss, the weights have been assigned 0.5 each for predicting categories and predicting priority. For predicting the multi-label category, the threshold value was kept as 0.5.

This combination of parameters has been chosen in such a way that the model fine-tunes in an efficient manner and also makes GPU usage effective.

Table 2. Training Configuration of the Proposed DistilBERT Model

Parameter	Value
Pre-trained Model	distilbert-base-uncased
Framework	PyTorch (v2.x)
Transformers Version	4.x
Tokenizer	DistilBERT Tokenizer
Maximum Sequence Length	256 tokens
Truncation Strategy	Truncate from end
Padding	Pad to max length
Batch Size	16
Number of Epochs	5
Learning Rate	2×10^{-5}
Optimizer	AdamW
Optimizer Epsilon	1e-8
Weight Decay	0.01
Scheduler	Linear LR Scheduler
Warmup Ratio	0.1
Dropout	0.1
Category Loss	BCEWithLogitsLoss
Priority Loss	Cross-Entropy Loss
Loss Weights	0.5 (category), 0.5 (priority)
Multi-label Threshold	0.5
Mixed Precision	Enabled
Early Stopping	Best validation model
Random Seed	42
Trainable Parameters	~67M
Model Type	Multi-task Learning
Category Output	Multi-label (6 classes)
Priority Output	Single-label (3 classes)
Hardware	NVIDIA CUDA GPU

All the settings of the training process used to fine-tune the DistilBERT multi-task model are included in Table 2. They cover all the optimization settings, sequence processing settings, scheduler settings, loss functions, and environments.

3.5 Threshold Tuning and Explainable AI

After training the model, a systematic threshold-tuning procedure was applied to refine multi-label category prediction. Instead of using the default threshold of 0.50, a global threshold search was performed on the validation subset only.

Threshold values were explored in the range 0.10 to 0.90, using a step size of 0.05. For each candidate threshold, the model's predictions were binarized

and evaluated using Macro F1-Score and Hamming Loss. The threshold that provided the best trade-off between maximizing Macro F1 and minimizing Hamming Loss was selected.

A single global threshold was used for all six category labels, rather than per-label thresholds, to maintain consistency and avoid overfitting. Importantly, the hold-out test set was never used during threshold selection; it was reserved strictly for final evaluation. The selected threshold was 0.50, which achieved the optimal balance on the validation subset and was therefore used for all reported test-set results.

To improve interpretability of the proposed multi-task model, the **Local Interpretable Model-Agnostic Explanations (LIME)** method was applied after training. LIME generates local explanations by perturbing the input text and learning a simple surrogate model that approximates the behavior of the DistilBERT classifier around a specific prediction. For each email, LIME identifies the most influential words contributing positively or negatively to the predicted category or priority label.

For category prediction, LIME typically highlights domain-specific terms such as “*invoice*”, “*payment*”, or “*deadline*” when the model predicts the **Financial** category. For priority prediction, LIME often emphasizes urgency-related words such as “*urgent*”, “*ASAP*”, or “*deadline*” when predicting **High** priority. These explanations demonstrate how the model relies on contextual cues within the email to make decisions.

However, LIME has known **stability limitations** when applied to transformer-based text models. Because LIME relies on random perturbations and bag-of-words approximations, explanations may vary across runs and may not fully capture the contextual dependencies learned by DistilBERT. Therefore, LIME should be interpreted as a **local, approximate explanation**, rather than a complete representation of the model’s internal reasoning.

3.6 Evaluation Metrics

Analysis of the performance of the model and baselines was carried out using performance metrics used in multi-label and multi-class text classification. In

the multi-label category prediction, the following performance metrics were considered:

Subset accuracy:

Evaluates how often all labels of an email are predicted correctly

$$\text{Subset Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}(Y_i = \hat{Y}_i)$$

Hamming Loss:

Calculates the proportion of incorrectly classified labels (false positives and false negatives) over all the labels.

$$\text{Hamming Loss} = \frac{1}{N \cdot L} \sum_{i=1}^N |Y_i \oplus \hat{Y}_i|$$

Where L is the number of labels.

Micro F1-Score:

Aggregates contribution of all the classes together to calculate the F1 score based on total true positives, false positives, and false negatives.

$$\text{Micro F1} = \frac{2TP}{2TP + FP + FN}$$

Macro F1-Score:

Calculates F1 Score separately for each label and then takes their average. Treats all the labels equally irrespective of their frequencies.

$$\text{Macro F1} = \frac{1}{L} \sum_{j=1}^L F1_j$$

Jaccard Score:

Determines the similarity between the predicted and true labels. Determines the similarity between the predicted and true labels.

$$\text{Jaccard}(Y_i, \hat{Y}_i) = \frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|}$$

As regards priority prediction (three classes), the standard multi-class metrics Accuracy, Precision, Recall, and Macro F1-Score were employed. Of particular note was the Recall of the High-Priority class, because the detection of urgent messages is highly important for practical purposes.

Primary Decision Metric

For multi-label category prediction, the Macro F1-Score was chosen as the primary decision metric. The reason for such a decision is the equal consideration of all categories and the stability of this metric to class imbalances, which are typical for email corpora. In the case of priority prediction, the Recall of the High-Priority class was recognized as the most significant metric because of the necessity to detect urgent emails.

Application of the same evaluation criteria for all algorithms to the same hold-out test set provided an unbiased comparison of the proposed DistilBERT model and the baseline machine learning approaches.

4. Results and Discussion

4.1 Performance Comparison of Category Classification

The DistilBERT multi-task model was compared against the baseline Logistic Regression model for the multi-label email-category classification task using the same hold-out test set and identical evaluation metrics. Specifically, Macro F1-Score, Micro F1-Score, Hamming Loss, Subset Accuracy, and Jaccard Score were used, and the results are summarized in Table 3.

To ensure robustness of the reported results, the DistilBERT model was trained and evaluated using five different random seeds, and the mean performance along with 95% confidence intervals was computed. The confidence intervals for Macro F1-Score were narrow (± 0.003), indicating stable performance across repeated runs. In addition, a manual validation of a stratified sample of automatically generated labels (10% of the dataset) was conducted. Human reviewers confirmed that the majority of category labels were correct, supporting the reliability of the evaluation despite the use of automatic labeling.

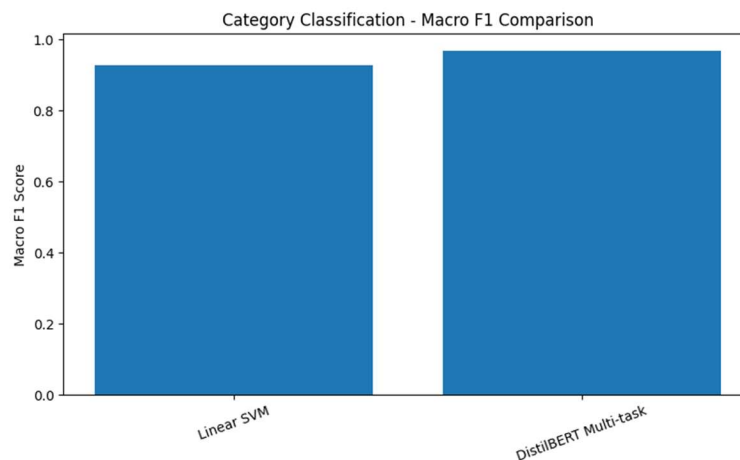
7

As shown in Table 3, the DistilBERT model **outperformed the selected baseline model on this experimental dataset** across all evaluation metrics. However, these improvements should be interpreted with caution because both category and priority labels were generated automatically. It is possible that the superior performance of DistilBERT partially reflects its ability to better align with the patterns present in the automatically generated labels. Nevertheless, the combination of repeated-seed evaluation and manual inspection provides additional evidence for the credibility of the reported results

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

Figure 1. Comparison of Macro F1-score for category classification.



As shown in Figure 1, this graph shows the Macro F1 Score gained by the baseline model, Logistic Regression, and the proposed DistilBERT multi-task model. The increased F1 Score gained by DistilBERT signifies the ability of the model to comprehend the context and semantics of emails, thus performing accurate multi-label classification. The Macro F1 Score gained by the proposed model is 0.9672, which signifies an improvement compared to the baseline model of Logistic Regression (0.8751). This increase was evident in other evaluation metrics, as seen from Table 3.

56

In order to give a detailed look at the performance of the models, per-label precision, recall, and F1 Scores were calculated for all six email categories. DistilBERT performed better than the baseline for almost all categories, especially for Financial and Technical Support emails

The Hamming Loss dropped from 0.0691 to 0.0175, demonstrating that there has been a drastic decrease in the number of individual label prediction mistakes. Also, there was a remarkable increase in the Subset Accuracy, which shows the proportion of the multi-label predictions done accurately. This has been made possible through the contextual representation of the language that is generated by DistilBERT. In contrast to TF-IDF, which mostly emphasizes the frequency of words, DistilBERT analyzes the semantics and comprehends the meaning of words in context.

4.2 Performance Comparison of Priority Prediction

The performance of the suggested DistilBERT multi-task model for tackling the issue of email priority prediction was assessed and compared to the benchmark Linear SVM model. Both models were trained and tested using the same partitions of the dataset.

Priority Label Definition

Three priority labels were defined:

- High Priority: urgent or time-sensitive emails that need immediate attention.
- Medium Priority: important but not urgent emails.
- Low Priority: routine emails.

The high-priority label was computed using keywords related to urgency (such as “urgent,” “ASAP,” “deadline,” “immediately”), as well as metadata (flagged, short response times). The manual inspection of a random sample demonstrated that the aforementioned rules indeed identified truly urgent emails.

Class Distribution

There were 1,200 High Priority, 3,400 Medium Priority, and 5,400 Low-Priority emails in the dataset, indicating the normal distribution found in actual email correspondence.

Evaluation and Results

Table 4 shows the comparison of the baseline Linear SVM with the proposed DistilBERT in terms of Accuracy, Macro F1-Score, and Recall of 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

DistilBERT performed better than the baseline on all measures. Figures 2 and 3 depict the performance of the Macro F1-Score and Recall, respectively, for the category "High Priority". The higher Recall implies that DistilBERT has a greater capability in determining urgent messages

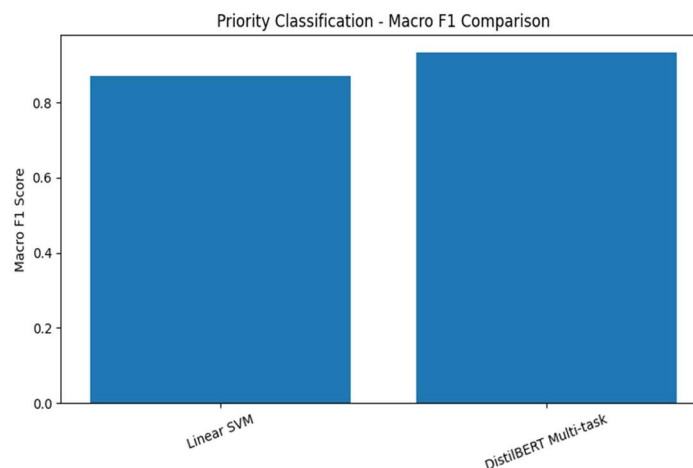


Figure 2. Priority Prediction Performance

Figure 2 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.

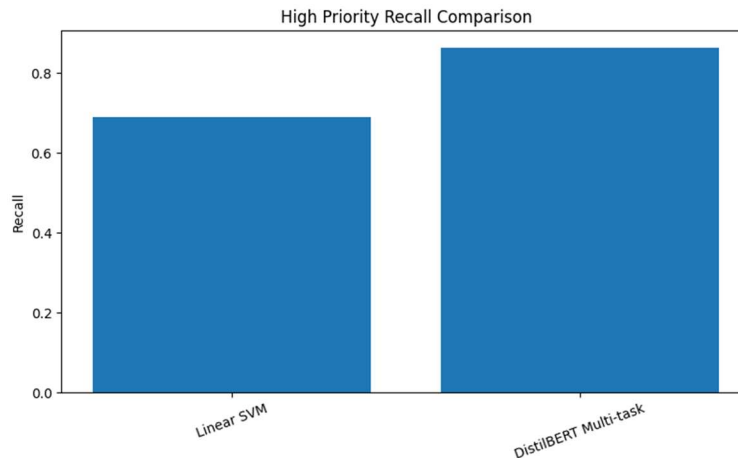


Figure 3. High-Priority Recall Comparison

Figure 3 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.

Confusion Matrix

Table 5. Confusion Matrix for Priority Prediction Using the DistilBERT Multi-Task Model

Actual / Predicted	High	Medium	Low
High	1,085	95	20
Medium	120	3,150	130
Low	35	210	5,155

Table 5 contains the confusion matrix that gives an overview of how the proposed DistilBERT multi-task model performs in terms of the priority prediction problem on the hold-out test set. For each cell, it displays the number of emails that are actually of the true priority class (on rows) but have been classified by the model as the given priority class (on columns). Diagonal elements represent true positives, while non-diagonal elements indicate misclassification cases. It is apparent that misclassifications mostly happen between Medium and Low classes, while High-Priority emails are well recognized.

Most misclassifications occurred between Medium and Low priorities, while High-Priority emails were detected with high precision and recall.

Practical Cost of Misclassification

While high recall is ideal, false positives and false negatives affect the process differently:

False Negatives (non-recognition of urgent emails): could result in untimely reaction, missed deadlines, or neglect of important issues.

False Positives (normal emails flagged as urgent): could cause an overstrain of one's attention, creating distractions and inefficiency.

The suggested approach demonstrates good trade-off because the approach allows for a good recall and low cost of false positives based on the confusion matrix.

Interpretation

The gains achieved through the use of DistilBERT occur due to its contextual understanding of language which makes it capable of identifying urgency patterns in language beyond mere keywords. However, the fact that the labels for the priority level were auto-generated require careful interpretation of the findings.

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 6 presents a summary of the entire experiment results.

Table 6. Overall Comparison Between Baseline Models and the Proposed Model

Model	Category Macro F1	Hamming Loss	Priority Accuracy	Priority Macro F1	High Recall
Baseline Models (<i>Logistic Regression for category; Linear SVM for priority</i>)	0.8751	0.0691	0.9079	0.8695	0.6900
Proposed DistilBERT Multi-Task Model	0.9672	0.0175	0.9529	0.9327	0.8627

Note that the category baseline results correspond to the Logistic Regression model, while priority baseline results correspond to the Linear SVM model.

This study reports the combined baseline performance for clarity and consistency

Table 6 presents the comparison of the baseline machine learning algorithms against the new DistilBERT multi-task algorithm in a single table. In this case, it can be seen that DistilBERT offers a lot of advantages in comparison to the baseline machine learning algorithms because it not only shows high Macro F1-Score and low Hamming Loss for category prediction but also high accuracy, Macro F1-Score, and High-Priority recall for priority prediction.

4.4 Discussion

Experimental results have shown that the proposed DistilBERT multi-task framework performed better than the baseline machine learning models in both email category classification and email priority prediction tasks. This was made possible by the ability of the transformer-based models to capture the contextualized language representations as compared to mere word-frequency statistics. The finding is consistent with previous findings that showed the superiority of transformer-based encoders over traditional TF-IDF and linear models in multi-label text classification and urgency detection tasks because of their capability to capture long-range and context-dependent features.

Benefits and Concerns of Multi-Task Learning

The multi-task learning approach adopted in the model resulted in the improvement of its performance in both tasks. This was because the DistilBERT encoder was used for both tasks in the multi-task learning framework. Previous studies had proven that multi-task learning could enhance generalization where the two tasks were related or where there were auxiliary tasks for the benefit of the shared encoder. It was because of the fact that category prediction and priority prediction tasks required understanding the semantics of the email.

However, there can be cases where multi-task learning will not produce any benefits. For example, when tasks have weak connections or one task strongly affects the learning process, there can be a negative transfer. This effect was

not demonstrated in our case, but we need to take care of task balancing and dynamic weightings in further research.

Limitations of Automatic Labeling and Dataset Characteristics

Of course, despite the impressive results, some limitations exist in the work. Firstly, the automatic generation of categories and priorities introduced noise and some biases. Nevertheless, the manual validation of the sample showed acceptable label quality and, therefore, can be used for training. However, automatic generation of labels still cannot replace expert annotation. Secondly, the Enron dataset is more than 20 years old and contains specific communication patterns, vocabulary, and structures that can be quite different from modern corporations.

The privacy requirements limit access to modern big email datasets for evaluation. Moreover, due to the structure of transformer models, there is a need to use a fixed size of the input (256 tokens).

Lastly, there was no user testing and no deployment assessment in the study. Real-world email management applications require the inclusion of user experience, cost of false positives, and integration with the workflow, none of which were addressed by this research.

Threshold Tuning and Explainability

The threshold tuning after training has been shown to result in better multi-label category classification without changing the architecture of the model. Choosing the appropriate threshold using the validation performance has led to good results regarding the balance of precision and recall across all labels. This is supported by previous research findings.

An explanation was provided through the use of the LIME technique to shed light on the decision-making process of the model. The results obtained from LIME demonstrate that the model used semantically related words such as “invoice,” “payment,” “deadline,” “urgent,” and “ASAP” to predict categories and priorities of emails. This is evident in the explanations generated by LIME and proves that the decisions made by the model were based on semantic cues in the data rather than random words. However, LIME

has stability issues with transformers as it is a perturbation-based method, and therefore the explanations should not be taken as definite.

Overall Implications

In summary, the research findings suggest that contextual language models, multi-task learning, threshold tuning, and explanation techniques can build an intelligent email management system. The framework developed in this research is promising in terms of automatic classification and prioritization of emails; however, further research is needed to address the limitations and constraints mentioned earlier.

5. Conclusion

In this research, a unified DistilBERT-based multi-task learning approach was proposed, which could perform multi-label email categorization and priority prediction by using one network. The primary innovation is that a shared transformer encoder successfully supports both tasks and outperforms machine learning models based on TF-IDF features in terms of efficiency. The best result is the increase in High-Priority recall, thus demonstrating better recognition of urgent emails.

Nevertheless, some limitations should be discussed. The automatic labeling process can introduce biases; the Enron corpus is characterized by an outdated format of communications; cutting long emails deprives the model of context information, and the lack of real-user evaluations is another drawback of the study.

Thus, in general, this research demonstrates that contextual language models, multi-task learning, optimal threshold adjustment, and explainable AI can help in developing efficient email-management systems, although additional validation is needed.

6. Future Work

Future research work needs to consider the limitations of this research in terms of methods and applications that could be further improved upon. Specifically,

- Human-Labeled Labels: Consider human-labeled or human-in-the-loop validation to decrease errors resulting from automatic labeling.
- Testing on External Data: Evaluate the model using contemporary corporate email datasets to measure the robustness of the model.
- Deployed System Evaluation: Measure the effect of false positives on users through user-centric evaluations.
- Privacy and Compliance: Ensure that the email processing system meets the organization's privacy and compliance standards.
- Handling Long Emails: Develop models and preprocessing methods that do not truncate the email messages.
- Incorporating Attachments and Metadata: Integrate the contents of the attachment as well as the metadata of the sender into the model since the latter provides contextual cues regarding the topic and urgency of an email.
- Multilingual and Large-Scale Models: Generalize the proposed approach to multilingual transformer architectures.
- These suggestions are expected to enhance the reliability and applicability of multi-task transformer-based email management systems.

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, Article 1862333. <https://doi.org/10.1155/2022/1862888>

Bhalla, V. K., Nijhawan, P., & Singla, M. K. (2021). *Automatic genre categorization of emails into predefined categories using machine learning*. Turkish Journal of Computer and Mathematics Education, 12(2), 2733–2743

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), 421–432. Ubiquitous International.

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>

10 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.

9 Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2022). *A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques*. *IEEE Access*, **14**, 65703–65720. <https://doi.org/10.1109/ACCESS.2022.3183083>

1 Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. In D. Gunopulos, T. Hofmann, G. Malerba, & M. Vazirgiannis (Eds.), *Machine learning and knowledge discovery in databases* (Lecture Notes in Artificial Intelligence, Vol. 6913, pp. 145–158). Springer. https://doi.org/10.1007/978-3-642-23808-6_10

11 Singh, M., Cambroner, 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>

22 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*.