Predictive Analysis and Email Categorization Using Large Language Models

Sabih Ahmed[*1], Junaid Mazhar Muhammad[1], Muhammad Farrukh Shahid[1], M. Hassan Tanveer[2], and Rabab Raza[3]
[1]Department of Artificial Intelligence and Data Science (FAST-NUCES Karachi, Pakistan)
[2]Department of Robotics and Mechatronics Engineering (Kennesaw State University, Marietta, GA, USA)
[3]Department of International Relations and Mass communication (Karachi University, Karachi, Pakistan)
[*]**Correspondence:** s.ahmed27@hotmail.com

Citation | Ahmed. S., Muhammad. J. M, Shahid. M. F, Tanveer. M. H, Raza. R., "Predictive Analysis and Email Categorization Using Large Language Models", IJIST, Special Issue. pp 272-282, Oct 2024
DOI | https://doi.org/10.33411/IJIST/1104
Received | Oct 15, 2024 **Revised |** Oct 18, 2024 **Accepted |** Oct 23, 2024 **Published |** Oct 29, 2024.

With the global rise in internet users, email communication has become an integral part of daily life. Categorizing emails based on their intent can significantly save time and boost productivity. While previous research has explored machine learning models, including neural networks, for intent classification, Large Language Models (LLMs) have yet to be applied to intent-based email categorization. In this study, a subset of 11,000 emails from the publicly available Enron dataset was used to train various LLMs, including Bidirectional Encoder Representations from Transformers (BERT), Distil BERT, XLNet, and Generative Pre-training Transformer (GPT-2) for intent classification. Among these models, Distil BERT achieved the highest accuracy at 82%, followed closely by BERT with 81%. This research demonstrates the potential of LLMs to accurately identify the intent of emails, providing a valuable tool for email classification and management.

Keywords: BERT; Distil BERT; GPT-2; Large Language Based Models; Transformers; XL-Net.

Introduction:

The rapid growth of internet users has led to a significant increase in email communication. On average, users receive approximately forty to fifty emails daily, with organizations exchanging hundreds of emails and messages (as highlighted by an email statistics report [1] on the importance of email in organizations). Email communication is crucial for almost every organization, particularly in sectors like airlines, government institutions, and finance, as illustrated in Figure 1. These organizations rely heavily on email for communication, and employees often spend a substantial amount of time searching for and managing emails. Understanding the underlying intent behind emails can greatly improve user efficiency, leading to enhanced productivity. Emails with unclear intent are often overlooked or ignored in the inbox, reducing their effectiveness. Categorizing emails based on their intent involves understanding the sender's purpose and the content of their message. This approach, rooted in identifying latent intentions within emails, can significantly streamline email management and improve productivity. Manually categorizing emails can be a tedious task, and important emails may get lost or ignored.

Organizations also place a high value on effective customer support, and understanding the context of emails provides valuable insights to improve service. While prior research has explored intent-based identification and classification in text using machine learning techniques, there has been limited work on intent-based email classification [2][3], with very few implementations involving Large Language Models (LLMs) for email categorization [4]. LLMs, trained on large datasets, offer enhanced learning capabilities compared to traditional deep learning models and remain relatively unexplored in intent-based email categorization.

This thesis aims to explore various models for email categorization, including baseline deep learning (DL) models, and leverage LLMs such as BERT [18] to classify emails based on intent. The analysis will be conducted using the widely available Enron dataset [5], which contains approximately 500,000 enterprise emails from the Enron corporation. The primary advantage of using LLMs for email categorization lies in their data-driven approach, allowing them to adapt to real-world scenarios where predefined email categories may not exist, and categories can be determined based on the input data.

Objectives:

The main objectives and contributions of this research are as follows:

- We introduce a method for categorizing emails based on the underlying purpose or intent behind the messages.
- We implement an approach that utilizes Large Language Models (LLMs), such as BERT, Distil BERT, and GPT-2, to automatically categorize emails based on their content and the sender's intent.
- We conduct extensive experiments on the Enron dataset and demonstrate that our method outperforms existing approaches for intent-based email categorization.

Novelty Statement:

This research focuses on intent-based email classification using Large Language Models (LLMs), a technique that has not been previously applied in this domain. The paper is organized as follows: Section 2 reviews related works, presenting an overview of relevant approaches; Section 3 provides detailed insights into the dataset and methodologies employed; Section 4 discusses the experiments conducted and the simulation results; and Section 5 concludes the paper, offering potential directions for future research.

Literature Review:

Various machine learning models have been employed for intent-based classification in different domains. In [6], a single-layer Convolutional Neural Network (CNN) combined with BERT was used to detect intent from text in the Airline Travel Information Systems (ATIS) and

a manually prepared Chinese dataset from the Yuetongbao customer service platform. Among the models tested, BERT with CNN achieved the highest accuracy of 98.5%. In [3], the Avocado corpus, an enterprise email dataset, was used to detect different types of intents through sentence-level intent-based identification using a Dynamic-Context Recurrent Neural Network (DCRNN). In [4], a joint BERT model was utilized for intent classification in natural language understanding, specifically on the ATIS dataset. This joint BERT model outperformed individual BERT models, achieving an accuracy of 97.9%. In [2], intent-based segmentation was applied to classify emails from two datasets, the Enron corpus and Gmail accounts, using SVM and Naive Bayes models. In [7], BERT, SciBERT-cased, and other large language models were used for intent and sentiment classification of in-text citations from multiple datasets, including the Citation Sentiment Corpus (CSC), the Association for Computing Machinery (ACM) library, and the SciCite dataset. Both BERT and SciBERT-cased achieved an accuracy of 88%.

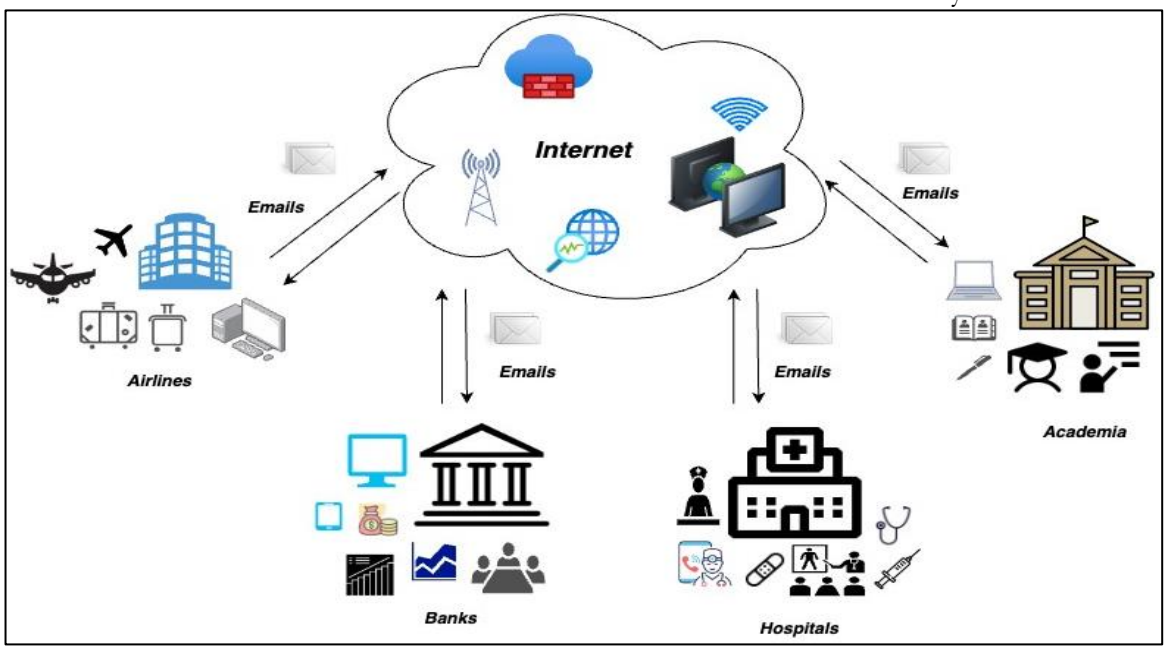

Figure 1. Email communication is crucial in industries such as banking, airlines, education, and healthcare.

In addition to intent-based classification, other types of email classification have been explored using various techniques. The paper [8] applied several machine learning algorithms, such as Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), Random Forest (RF), Decision Trees (DT), and Multi-Layer Perceptron (MLP), along with bio-inspired techniques like Particle Swarm Optimization (PSO) and Genetic Algorithm (GA), on datasets such as Enron, Spam Assassin, Ling-Spam, and PUA. The results showed that GA performed better with RF and DT models. In [9], different deep learning architectures combined with email representation techniques and word embeddings were used on the Ling spam, PU, Enron, and Apache Spam Assassin datasets. The Fast Text model combined with CNN and LSTM achieved the highest accuracy of 95.9%. In [10], a hybrid CNN-LSTM model was applied for spam SMS classification, achieving an accuracy of 98.37%, outperforming all other models tested.

Large language models (LLMs) have also been utilized for email classification. The paper [11] employed the BERT model for spam email detection, using 12 encoders from transformers and datasets like Spam Assassin, SMS Spam Collection, Ling-Spam, and Enron. Their proposed model achieved impressive F1-scores of 97.83%, 99.28%, 99.13%, and 98.62% on these datasets. In [12], LLMs were used on multiple datasets, including Enron, Spam Assassin, Ling-Spam, and SMS Spam Collection, for predictive analysis. The study compared models such as

RoBERTa, Set Fit, and Spam-T5, with Spam-T5 achieving the highest F1-score of 0.7498. In [13], BERT was applied to classify text on the UCI email and BBC News datasets, achieving accuracies of 91% and 89%, respectively. The paper [14] utilized BERT on multiple datasets, including Enron, IMDb, IMDb62, and Blog, for author classification. In [15], Distil BERT and BERT were applied to classify online news related to Covid-19, with Distil BERT achieving the highest accuracy of 95%. Inspired by these previous studies, which used deep learning models for various email classification tasks, this research proposes a supervised approach using large language models for intent-based email categorization. Table 1 provides a summary of the existing works in this area.

Table 1. Overview of existing works

References	Model(s) & Accuracy	Objective	Dataset
[6]	CNN (74.7%), BERT-LSTM (96.31%), BERT-GRU (96.86%), BERT-CNN (98.5%),	To compare performance between different models for intent based classification	ATIS (Airline Travel Information Systems) dataset and Chinese dataset
[3]	DCRNN (schedule meeting task f1measure: 73.48, promise action task f1-measure: 80.42, request information task f1-measure: 78.37) Accuracy not reported	To detect different types of intents in emails conversations	Avocado corpus
[4]	Joint BERT (97.5%), Joint BERT with conditional random field (97.9%)	To experiment BERT pre-trained model for joint intent classification and natural language understanding (NLU)	ATIS dataset
[2]	Naive bayes (89%) and SVM (90%)	To experiment intention-based segmentation to classify emails	Enron dataset and emails from gmail account
[7]	BERT (88%) and Sci BERT-cased (88%)	To classify in-text citations on the basis of intent and sentiments	Citation Sentiment Corpus (CSC), Association for Computing Machinery (ACM) library and Sci Cite dataset
[8]	SGD (97.64%), MNB (98.47%), DT (92,28%), RF (90.81%), ML (97.18%)	To perform comparative analysis for spam detection with bio-inspired models	PUA, Enron, Spam Assasin, and Ling-Spam
[9]	Word embedding with CNN and LSTM (95.59%), Fast Text with CNN and LSTM (95.9%), Keras embedding with CNN and LSTM (94.8%), and Keras embedding with CNN (94.5%)	To compare DeepLing Learning models with word embedding for classification of spam emails	spam, PU, Enron and Spam Assassin

Continued on next page

Material and Methods:

Dataset: The Enron dataset [5], available online, contains approximately 500,000 emails. This dataset comprises enterprise emails from Enron Corporation, including full email bodies,

headers, and other associated metadata. For this research, a subset of 11,098 emails was used. As no publicly available datasets contained intention-based labels, the emails were manually categorized into specific intent categories. Each email was labeled according to the sender's purpose, with the intent categories including "inform," "deliver," "request," "query," and "remind." The distribution of these categories is illustrated in Figure 2.

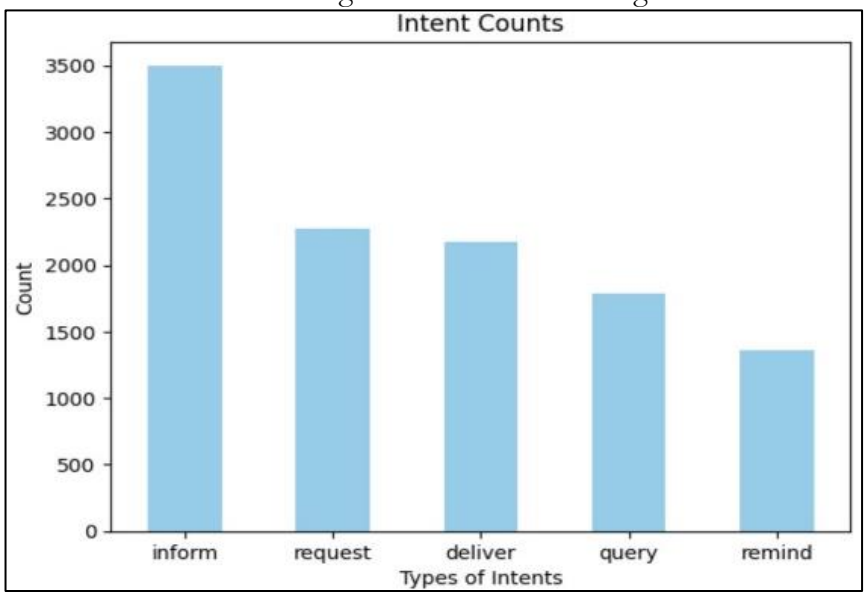

Figure 2. Labelled Enron email categories distribution

Table 2. Overview of existing works (continued)

References	Model(s) & Accuracy	Objective	Dataset
[10]	SVM (97.8%), KNN (90.0%) DT (96.5%), LR (96.5%), RF (97.8%), Ada Boost (97.2%), Bagging classifier (97.2%), Extra trees (97.8%), CNN (98.1%), LSTM (98.1%), CNNLSTM (98.3%))	To perform comparative analysis for different classification models and find best approach to detect spam sms	SMS Spam and set of Arabic messages (from local phone)
[11]	BERT (Enron f1-measure:98.62%, Spam Assasin f1-measure:97.83%, Ling-Spam f1-measure:99.13% and SMS spam collection f1 measure:99.28%) Accuracy not reported	To further improve spam prediction using BERT based model	Enron corpus, Spam Assassin corpus, Ling-Spam corpus and SMS spam collection corpus
[13]	BERT (UCI email dataset: 91.0%, BBC News dataset: 89%),	To experiment fine-tuned BERT model for text document classifications	UCI email dataset and BBC News dataset
[14]	BERT (Enron dataset: 99.95%, IMDb dataset: 99.6%, Blog dataset: 61.3%),	To explore fine-tune a pre-trained BERT model for author classification experiments	Enron, IMDb and Blog datasets

[15]	BERT (balanced dataset: 97%, imbalanced dataset: 96%), Distil BERT (balanced dataset: 95%, imbalanced dataset: 94%),	To explore online news binary classification on Covid19 online news data using Distil BERT and BERT models	News web- sites (10news.com, cnn.com, and foxla.com)
Our contribution	BERT (81.0%), Distil BERT (82.0%), XL Net XL Nete Xtreme Language understanding Network (78%), GPT-2 (71%)	Intent based classification using Large Language Models	Enron dataset

Proposed Methodology:

This section outlines the methodology steps employed in the development of the proposed framework:

Preliminary Overview of LLMs:

LLMs are advanced artificial intelligence models built using deep learning techniques, particularly the transformer architecture, which allows them to identify intricate patterns and relationships within large datasets [16]. The transformer architecture serves as the foundational structure for all LLMs. Key components of the transformer architecture, as described in [17], are as follows:

Embedding:

This step occurs at the lower part of the encoder, where each word in the sentence is converted into a vector within a high-dimensional vector space. It is responsible for capturing the semantic meaning of words and phrases by transforming them into numerical representations.

Self-Attention Mechanism:

This step captures the contextual meaning of the input sequence. Here, a score is computed to determine the amount of attention that should be allocated to the surrounding context while encoding a word at a specific position.

Masked Multi-headed Attention:

This step in the decoder portion masks the output by setting the probabilities of the masked values to zero, ensuring they are not considered during processing.

BERT:

BERT [18] is a bidirectional transformer model trained on a large dataset of approximately 3.3 billion words sourced from Wikipedia and a corpus containing Google Books data. During its training, BERT employs two key techniques: Masked Language Modeling (MLM) and Next-Sentence Prediction (NSP). In the MLM phase, certain words in a sentence are masked, prompting the model to predict the missing tokens by processing the sentence bidirectionally. In the NSP phase, the model learns to identify the relationship between two consecutive sentences, determining whether the second sentence logically follows the first.

Distil BERT:

Distil BERT [19] shares the same general architecture as BERT [18], but it is a smaller, more efficient version. Distil BERT is 40% smaller than BERT, yet it retains approximately 97% of BERT's language understanding capabilities while operating 60% faster.

XL Net:

XL Net [20] is an eXtreme Language understanding Network that employs a generalized autoregressive pre-training method. Unlike traditional bidirectional language models, XL Net considers all possible permutations of word order in a sentence. This approach enables the model to learn a more comprehensive understanding of context, rather than simply predicting the next word in a fixed sequence.

GPT2:

GPT-2 [21] is a large transformer model with 1.5 billion parameters, trained on an extensive Web Text dataset containing 8 million web pages. The architecture of GPT-2 consists of multiple transformer decoder layers, each featuring self-attention mechanisms and feed-forward neural networks. Key components of this model include the self-attention mechanism for capturing contextual relationships, feed-forward neural networks to identify complex patterns in text, and layer normalization with residual connections to support efficient learning.

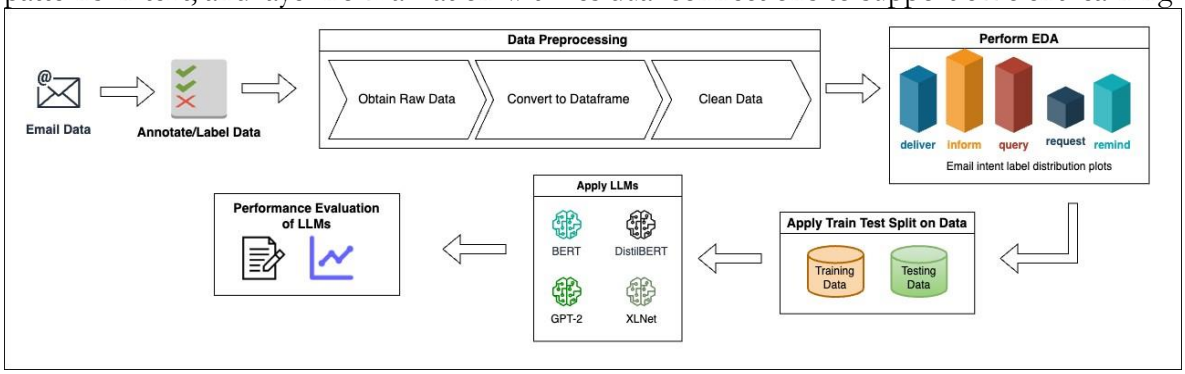

Figure 3. Proposed work methodology showing the LLM implementation with necessary pre-processing steps.

Methodology:

Figure 3 illustrates the methodology used in this study. First, 11,098 emails from the Enron dataset were collected, followed by pre-processing steps such as handling null values. The emails were then labeled into the following intent categories: inform, remind, request, query, and deliver. After labeling, the dataset was divided into training (70%) and testing (30%) sets using stratified splitting to ensure a consistent distribution of email intents across both sets. A 70:30 split was chosen due to its computational efficiency compared to more resource-intensive methods like k-fold cross-validation, which would require multiple rounds of training and validation. This approach is particularly advantageous for LLMs like BERT and Distil BERT, as it saves time and resources while maintaining strong performance. The models implemented in this study include BERT, Distil BERT, GPT-2, and XL Net. These models were fine-tuned to classify the emails according to their labeled intents. Finally, the performance of these models was assessed using the test data, with metrics such as accuracy, precision, recall, and F1-score employed to evaluate each model's effectiveness in intent classification.

Results:

For all experiments, various numbers of epochs were tested to determine the optimal settings. The dataset was split into training, validation, and test sets to ensure an accurate performance evaluation. Figure 4 presents the overall experiment results, including the Train vs. Validation Loss and Train vs. Validation Accuracy for the LLMs used (Distil BERT, BERT, XL Net, and GPT-2).

BERT:

The BERT pre-trained base uncased model was trained for a total of 5 epochs with a batch size of 16 and a learning rate of 1e-5. This pre-trained model architecture consists of 12 transformer layers, each with a hidden size of 768. Additionally, it includes a linear layer and a SoftMax function for generating the output. Figure 4.c illustrates the training loss vs. validation loss over the epochs. In the first epoch, the validation loss showed a significant decrease; however, the rate of reduction slowed in subsequent epochs, while the training loss consistently decreased. Figure 4.d shows the trend of training and validation accuracies, where the training accuracy steadily increased, and the validation accuracy gradually improved in the first epoch before plateauing. Table 3 presents the classification report, including accuracy, precision, recall, and F1-score using the BERT model.

Table 3. BERT Classification Report

	Precision	Recall	F1-score	Support
Deliver	0.76	0.85	0.80	327
Inform	0.80	0.79	0.80	525
Query	0.84	0.88	0.86	268
Remind	0.85	0.78	0.81	204
Request	0.81	0.76	0.79	341
Accuracy value			0.81	1665
Macro value avg	0.81	0.81	0.81	1665
Weighted avg value	0.81	0.81	0.81	1665

Distil BERT:

The Distil BERT model, based on the pre-trained Distil BERT base uncased architecture, was trained for 5 epochs with a batch size of 16 and a learning rate of 1e-5. Figure 4.a shows the trends of training loss versus validation loss over the epochs. The training loss decreased steadily, dropping to a very low value. In the first epoch, the validation loss decreased steadily, but in subsequent epochs, the rate of decrease remained constant. Figure 4.b illustrates the trends between training and validation accuracies. Both the training and validation accuracies showed consistent improvement in the first epoch, after which the validation accuracy plateaued. Table 4 presents the classification report, which includes accuracy, precision, recall, and F1-score for the Distil BERT model.

Table 4. Distil BERT Classification Report

	Precision	Recall	F1-score	Support
Deliver	0.69	0.87	0.77	237
Inform	0.87	0.75	0.80	339
Query	0.86	0.90	0.88	187
Remind	0.91	0.85	0.88	114
Request	0.88	0.81	0.84	234
Accuracy value			0.82	1111
Macro value avg	0.84	0.84	0.83	1111
Weighted avg value	0.83	0.82	0.82	1111

Table 5. XL Net Classification Report

	Precision	Recall	F1-score	Support
Deliver	0.84	0.69	0.76	653
Inform	0.77	0.76	0.76	1050
Query	0.83	0.86	0.85	537
Remind	0.90	0.71	0.79	408
Request	0.66	0.85	0.74	682
Accuracy value			0.78	3330
Macro value avg	0.80	0.77	0.78	3330
Weighted avg value	0.79	0.78	0.78	3330
Weighted avg value	0.79	0.78	0.78	3330

XL Net:

The XL Net model, based on the pre-trained XL Net base cased architecture, was trained for 3 epochs with a batch size of 16 and a learning rate of 3e-5. Figure 4.e displays the comparison of training and validation loss trends. During the first epoch, the training loss decreased sharply, followed by a steady decline in the next two epochs. In contrast, the validation loss showed a gradual decrease in the first epoch, but began to increase in the subsequent epochs. Figure 4.f illustrates the trends in training and validation accuracies. Table 5 presents the classification report, including accuracy, precision, recall, and F1-score for the XL Net

model.

GPT-2:

The GPT-2 model was trained for 3 epochs with a batch size of 16 and a learning rate of 5e-5, using the pre-trained GPT-2 architecture. Figure 4.g shows the trends in training loss versus validation loss throughout the epochs. The training loss decreased consistently, while the validation loss followed a steady trend during the first epoch, but began to increase in the subsequent epochs. Figure 4.h illustrates the trends in training and validation accuracies. Table 6 presents the classification report, including accuracy, precision, recall, and F1-score for the GPT-2 model.

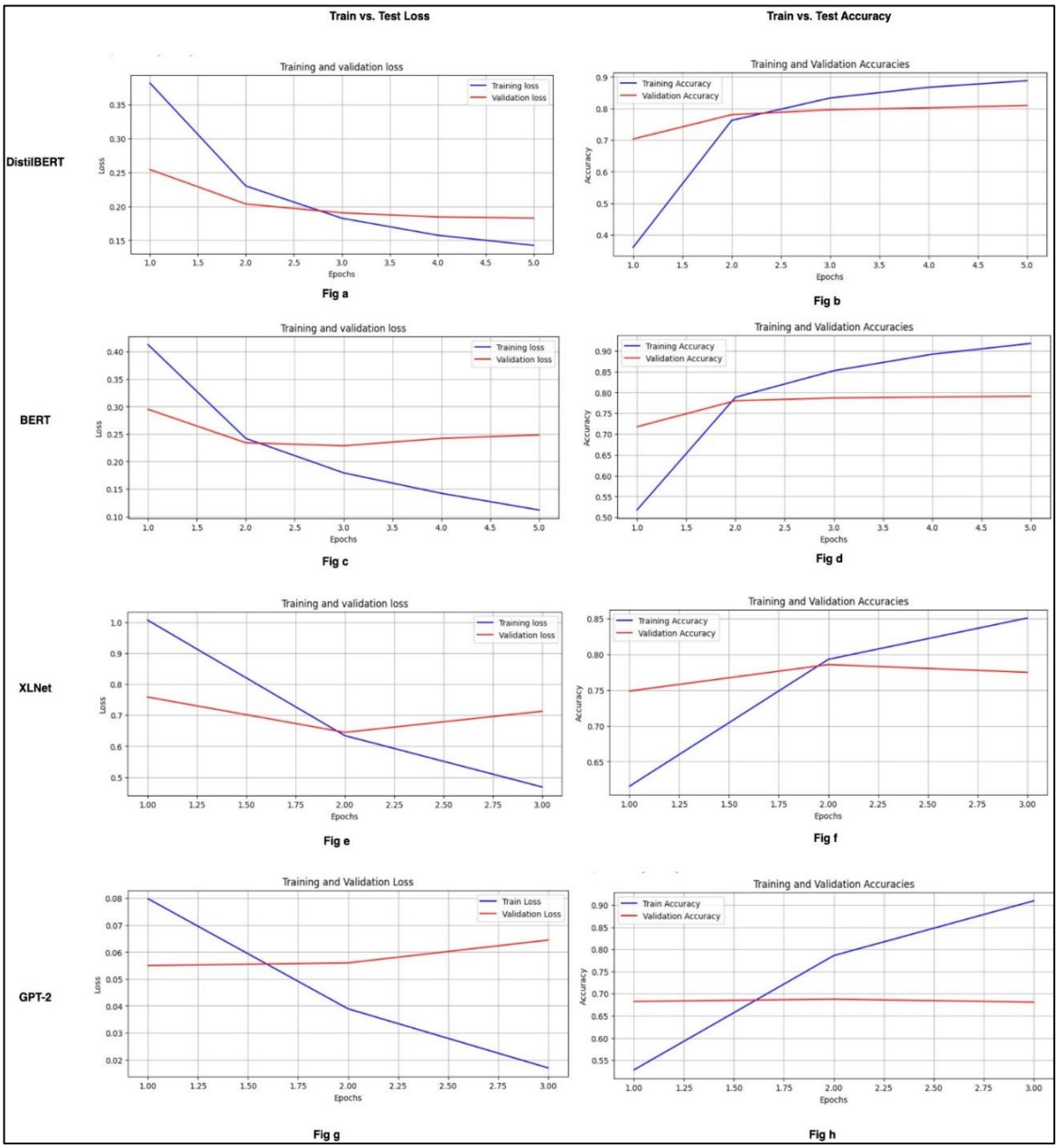

Figure 4. Results comparison between LLM

Discussion:

Among the evaluated LLMs, Distil BERT achieved the highest accuracy of 82%, making it the top-performing model in this comparison. BERT closely followed with an accuracy of 81%, demonstrating strong performance as well. Table 7 summarizes the results obtained with each LLM.

Table 6. GPT-2 Classification Report

	Precision	Recall	F1-score	Support
Inform	0.67	0.75	0.70	1050
Deliver	0.68	0.67	0.67	653
Remind	0.77	0.64	0.69	408
Request	0.73	0.69	0.71	682
Query	0.78	0.75	0.76	537
Accuracy value			0.71	3330
Macro value avg	0.72	0.69	0.71	3330
Weighted avg value	0.71	0.70	0.71	3330

Table 7. Summary of Results

Model	Accuracy
Distil BERT	0.82
BERT	0.81
XL Net	0.78
GPT-2	0.71

Conclusion:

This research utilized LLMs such as BERT, GPT-2, Distil BERT, and XL Net to classify emails based on the sender's intent, a novel approach within this domain. The dataset used was a subset of the widely studied Enron dataset, which includes real-life emails. Approximately 11,000 emails were labeled with intents such as inform, remind, request, query, and deliver. The accuracy achieved by each model was 81.0% for BERT, 82.0% for Distil BERT, 78.0% for XL Net, and 71% for GPT-2, using the intent-labeled data. This research effectively demonstrates the potential of LLMs to accurately understand email intent. Future work could involve automating the process to intelligently categorize emails based on intent. For example, if the support team is responsible for handling emails related to "request access," an intelligent categorization system could automatically route these emails to the appropriate team, improving efficiency and response times within organizations.

Acknowledgement: The authors want to acknowledge FAST-NUCES for providing the facility for conducting this research.

Author's Contribution: Conceptualization, Muhammad Shahid; Formal analysis, Muhammad Shahid and M. Hassan, Sabih Ahmed and Junaid Muhammad; Data Labelling, Sabih Ahmed and Rabab Raza; Methodology, Sabih Ahmed; Validation, Muhammad Shahid; Writing-original draft, Sabih Ahmed and Junaid Muhammad; Writing - review editing, M. Hassan.

Conflict of Interest: The authors report no conflicts of interest.

References:

[1] S. Radicati and J. Levenstein, "Email statistics report, 2021-2025," Radicati Group, Palo Alto, CA, USA, Tech. Rep, 2021.

[2] S. K. Sonbhadra, S. Agarwal, M. Syafrullah, and K. Adiyarta, "Email classification via intention-based segmentation," in 2020 7th International Conference on Electrical Engineering, Computer Sciences and Informatics (EECSI), pp. 38– 44, IEEE, 2020.

[3] W. Wang, S. Hosseini, A. H. Awadallah, P. N. Bennett, and C. Quirk, "Context-aware intent identification in email conversations," in Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 585–594, 2019.

[4] Q. Chen, Z. Zhuo, and W. Wang, "Bert for joint intent classification and slot filling," arXiv preprint arXiv:1902.10909, 2019.

[5] "Enron Dataset." https://www.cs.cmu.edu/~./enron/, 2015. Accessed: 2023-09-07.

[6] C. He, S. Chen, S. Huang, J. Zhang, and X. Song, "Using convolutional neural network with bert for intent determination," in 2019 International Conference on Asian Language Processing (IALP), pp. 65–70, IEEE, 2019.

[7] R. Visser and M. Dunaiski, "Sentiment and intent classification of in-text citations using bert," in Proceedings of 43rd Conference of the South African Insti, vol. 85, pp. 129–145, 2022.

[8] S. Gibson, B. Issac, L. Zhang, and S. M. Jacob, "Detecting spam email with machine learning optimized with bioinspired metaheuristic algorithms," IEEE Access, vol. 8, pp. 187914–187932, 2020.

[9] S. Srinivasan, V. Ravi, M. Alazab, S. Ketha, A. M. Al-Zoubi, and S. Kotti Padannayil, "Spam emails detection based on distributed word embedding with deep learning," Machine intelligence and big data analytics for cybersecurity applications, pp. 161–189, 2021.

[10] A. Ghourabi, M. A. Mahmood, and Q. M. Alzubi, "A hybrid cnn-lstm model for sms spam detection in arabic and english messages," Future Internet, vol. 12, no. 9, p. 156, 2020.

[11] T. Sahmoud and D. M. Mikki, "Spam detection using bert," arXiv preprint arXiv:2206.02443, 2022.

[12] M. Labonne and S. Moran, "Spam-t5: Benchmarking large language models for few-shot email spam detection," arXiv preprint arXiv:2304.01238, 2023.

[13] M. A. Shah, M. J. Iqbal, N. Noreen, and I. Ahmed, "An automated text document classification framework using bert," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, 2023.

[14] M. Fabien, E. Villatoro-Tello, P. Motlicek, and S. Parida, "Bertaa: Bert fine-tuning for authorship attribution," in Proceedings of the 17th International Conference on Natural Language Processing (ICON), pp. 127–137, 2020.

[15] S. K. Akpatsa, H. Lei, X. Li, V.-H. K. S. Obeng, E.M. Martey, P.C. Addo, and D. D. Fiawoo, "Online news sentiment classification using distil bert.," Journal of Quantum Computing, vol. 4, no. 1, 2022.

[16] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al., "Training compute-optimal large language models," arXiv preprint arXiv:2203.15556, 2022.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[19] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[20] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," Advances in neural information processing systems, vol. 32, 2019.

[21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., "Language models are unsupervised multitask learners," OpenAI blog, vol. 1, no. 8, p. 9, 2019.