NLP-Project "Automatic Summarization of Email Threads Using Natural Language Processing"

Aly Ahmed El-Azazy & Mohamed Samy Moussa

1 Introduction and Motivation

Email communication remains one of the most widely used forms of information exchange in professional, academic, and organizational environments. As discussions evolve over time, email conversations often grow into long and complex threads involving multiple participants, forwarded messages, and repeated replies. While email threads support collaboration and record-keeping, their increasing length and redundancy make it difficult for users to quickly identify essential information such as key decisions, updates, or required actions.

The rapid growth of digital communication has intensified the problem of information overload. In many real-world scenarios, users are required to scan extensive email threads in order to extract a small amount of relevant information. This manual process is time-consuming, error-prone, and inefficient, particularly in fast-paced organizational settings where timely understanding of information is critical. As a result, there is a growing need for automated tools that can condense lengthy email conversations into concise and informative summaries.

Natural Language Processing (NLP) offers effective techniques to address this challenge through automatic text summarization. Text summarization aims to generate shorter representations of longer texts while preserving their core meaning. In the context of email communication, summarization can significantly reduce cognitive load by allowing users to grasp the main points of an email thread without reading every individual message. However, summarizing email threads presents unique challenges compared to traditional single-document summarization, as email threads consist of multiple messages written by different authors over time and often contain informal language, repeated metadata, forwarded content, and formatting noise.

The motivation of this project is to study automatic summarization of email threads using a real-world dataset that includes both raw email content and corresponding human-generated summaries. The availability of reference summaries enables meaningful analysis of summarization quality and provides valuable insight into how important information is distilled from complex conversational data. By analyzing the structure and characteristics of email threads,

this project aims to highlight the challenges associated with conversational summarization and to establish a strong foundation for further exploration of NLP-based summarization approaches.

2 Literature Review

Text summarization is a fundamental task in Natural Language Processing (NLP) that aims to automatically generate concise and informative representations of longer documents. With the rapid growth of digital communication such as emails, reports, and online content, summarization has become increasingly important for addressing information overload. Early summarization techniques relied heavily on statistical and graph-based methods, such as TextRank [1], which rank sentences based on lexical similarity and structural importance. While these approaches are computationally efficient and preserve grammatical correctness, they are limited in their ability to capture deeper semantic relationships and often struggle with redundancy, particularly in conversational and multi-document settings. Recent survey studies highlight a clear shift from these traditional approaches toward deep learning and transformer-based models, emphasizing the need to handle diverse document structures, domain-specific language, and varying document lengths [2].

Extractive summarization approaches generate summaries by selecting the most relevant sentences directly from the source text. Classical extractive methods typically rely on features such as term frequency, sentence position, and similarity measures, which makes them robust and easy to interpret. More recent extractive systems incorporate neural representations to improve sentence selection. Hassan et al. [3] proposed an attention-based neural extractive model that captures contextual dependencies between sentences, demonstrating improved performance over traditional extractive baselines, especially for longer documents. Despite these improvements, extractive summarization remains limited in handling redundancy and discourse coherence, as selected sentences may repeat similar information or lack a logical narrative flow. These limitations are particularly evident in email threads, where quoted replies and forwarded content introduce substantial repetition.

Abstractive summarization aims to generate novel sentences that paraphrase and condense the core meaning of the source text, thereby more closely resembling human-written summaries. Advances in neural sequence-to-sequence learning and transformer architectures have significantly improved abstractive summarization quality. Transformer-based models such as BART [4], T5 [5], and PEGASUS [6] have achieved state-of-the-art results across multiple benchmark datasets. Shakil et al. [7] provide a comprehensive survey of modern abstractive summarization techniques, highlighting both their strong performance and persistent challenges, particularly hallucination, where generated summaries include information not present in the original text. This issue is especially problematic in email communication, where factual accuracy and faithful representation of decisions or actions are critical.

Transformer-based models have become the dominant paradigm in modern summarization due to their ability to model long-range dependencies using self-attention mechanisms. Comparative studies show that model performance varies significantly depending on dataset characteristics and domain. Daraghmi [8] evaluated BART, T5, and PEGASUS across diverse datasets and found that conversational and multi-document summarization tasks remain more challenging than single-document summarization. Similarly, Dharrao et al. [9] reported strong results for transformer models on business news summarization while noting limitations related to contextual completeness and factual consistency. Recent work has also explored dialogue-aware transformer architectures that explicitly model speaker turns and conversational flow, such as DialogLM, which demonstrates improved performance on dialogue and multi-turn summarization tasks by leveraging long-context pretraining [10].

Email thread summarization is commonly categorized as a form of multi-document and conversational summarization, where the input consists of multiple interrelated messages authored by different participants over time. Zhang et al. [11] introduced EmailSum, one of the first large-scale datasets specifically designed for abstractive email thread summarization, highlighting the unique challenges posed by email data, including redundancy, quoted replies, and informal language. Compared to single-document summarization, email summarization must address temporal ordering, implicit context, and structural noise introduced by signatures and metadata. These characteristics significantly complicate summarization and reduce the effectiveness of methods designed for cleaner, well-structured documents [2].

Evaluating summarization quality remains an open research problem. Automated metrics such as ROUGE [12] are widely used due to their scalability and ease of comparison with human-written references, but they often fail to capture semantic equivalence, coherence, and factual correctness, particularly for abstractive and conversational summaries. Recent work has proposed embedding-based metrics such as BERTScore [13], which better capture semantic similarity between generated and reference summaries. Nevertheless, several studies emphasize that automatic metrics alone are insufficient and should be complemented with human evaluation, especially in email and conversational summarization settings where multiple valid summaries may exist [7, 8]. Overall, the literature demonstrates substantial progress in text summarization through deep learning and transformer-based approaches, while consistently highlighting redundancy, noise, factual reliability, and evaluation subjectivity as persistent challenges, motivating further analysis of real-world email thread datasets.

3 Dataset Analysis, Insights, and Limitations

3.1 Dataset Description

This project uses the Email Thread Summary Dataset, which consists of two primary components: a collection of email thread details and a corresponding

set of human-generated summaries. The dataset contains a total of 4,167 email threads and 21,684 individual emails, all written in English. Each email thread is identified by a unique thread identifier that links multiple messages belonging to the same conversation.

The email thread details include metadata such as subject, timestamp, sender, recipients, and the body of each email message. The summary file provides a concise human-written summary for each thread, offering a high-level overview of the conversation. The availability of aligned input text and reference summaries makes this dataset particularly suitable for studying automatic text summarization in a real-world conversational setting.

3.2 Exploratory Data Analysis

An exploratory analysis of the dataset reveals significant variation in the structure and length of email threads. Some threads consist of only a small number of short emails, while others contain long chains of messages with extensive quoted replies and forwarded content. This variation highlights the diversity of communication patterns present in real-world email data.

The body of the emails often includes repeated headers, forwarding markers, and formatting artifacts introduced during email transmission. These elements increase redundancy and noise within the dataset. Additionally, emails frequently contain long recipient lists and administrative metadata that do not contribute directly to the semantic content of the conversation. As a result, the proportion of informative text relative to total text length varies considerably across threads.

The human-generated summaries are generally much shorter than the original email threads and focus on conveying key updates, decisions, or actions discussed in the conversation. This contrast between lengthy input threads and concise summaries demonstrates the practical need for effective summarization techniques.

3.3 Insights from the Dataset

Analysis of the dataset provides several important insights relevant to the summarization task. First, email threads are inherently conversational and multi-document in nature, requiring summarization systems to integrate information from multiple messages written by different authors over time. Important details may appear in only one message, while other content may be repeated across several replies.

Second, the presence of redundant and noisy text suggests that summarization models must be robust to irrelevant information. Forwarded messages and quoted replies often restate earlier content, which can mislead extractive summarization methods into selecting redundant sentences. Abstractive approaches may offer greater flexibility but must maintain factual consistency to avoid generating incorrect information.

Finally, the dataset illustrates that many email threads revolve around administrative updates or information dissemination, where the core message can often be summarized in one or two sentences. This characteristic aligns well with the goal of producing concise summaries that reduce cognitive load for users.

3.4 Dataset Limitations

Despite its strengths, the dataset has several limitations that must be considered. One major limitation is the presence of significant noise in the email bodies, including formatting artifacts, email signatures, and repeated headers. These elements can negatively affect model performance if not properly addressed during preprocessing.

Another limitation is the informal and domain-specific nature of email communication. The language used in emails may vary widely depending on organizational context, making it challenging to generalize summarization models trained on this dataset to other domains. Additionally, the meaning of certain messages may depend heavily on prior context or shared knowledge among participants, which may not be fully captured in the text alone.

Finally, while the dataset provides one human-generated summary per thread, summarization can be subjective, and alternative summaries may also be valid. This subjectivity introduces challenges for evaluation, as automated metrics may not fully capture summary quality or relevance.

3.5 Discussion

Overall, the Email Thread Summary Dataset offers a realistic and challenging testbed for studying conversational text summarization. Its combination of noisy real-world data and human-written summaries provides valuable opportunities for analyzing summarization methods and their limitations. Understanding the characteristics and constraints of the data set is essential to develop effective summarization models and to accurately interpret the experimental results.

References

[1] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2004.

[2] S. Khalid *et al.*, "Current trends and advances in extractive text summarization: A comprehensive review," *ResearchGate*, 2025, survey of modern extractive summarization techniques.

[3] A. Q. A. Hassan, B. B. Al-Onazi, M. Maashi, A. A. Darem, I. Abunadi, and A. Mahmud, "Enhancing extractive text summarization using an attention-based nlp model," *AIMS Mathematics*, vol. 9, no. 3, pp. 4616–4634, 2024.

[4] M. Lewis *et al.*, "Bart: Denoising sequence-to-sequence pre-training for natural language generation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 787–801, 2020.

[5] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[6] J. Zhang *et al.*, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *Proceedings of the International Conference on Machine Learning*, 2020.

[7] H. Shakil, A. Farooq, and J. Kalita, "Abstractive text summarization: State of the art, challenges, and improvements," *arXiv preprint arXiv:2409.02413*, 2024.

[8] E. Daraghmi, "A comparative study of pegasus, bart, and t5 for text summarization across diverse datasets," *Future Internet*, vol. 17, no. 9, p. 389, 2025.

[9] D. Dharrao, M. Mishra, A. Kazi, M. Pangavhane, and P. Pise, "Summarizing business news: Evaluating bart, t5, and pegasus for effective information extraction," *Revue d'Intelligence Artificielle*, vol. 38, no. 1, pp. 121–132, 2024.

[10] M. Zhong, D. Wang, and X. Qiu, "Dialoglm: Pre-trained dialogue transformer for long dialogue summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

[11] Y. Zhang, J. Li, M. Zhao, and Z. Liu, "Emailsum: Abstractive email thread summarization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[12] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004.

[13] T. Zhang, V. Kishore, F. Wu, K. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.