# **Literature Survey**

DOI: <https://ieeexplore-ieee-org.rlib.pace.edu/document/5170600>

Paper Title: **Email Grouping and Summarization: An Unsupervised Learning Technique**

Technique: Check the subject and sentences (phrases) to determine the email context.

Parent Email Grouping technique:

1. Find most frequently used words in the whole email (FW)
2. Find the activities (AC)
3. Group the activities with frequent words which is AC with FW to give AFW
4. Display the list of activities

Mapping the emails:

1. Record the information of sender, receiver, subject, time and the list of activities within the subject
2. If the email has the relevant wrt saved information with same sender, receiver and activities inside the email, they tend to fall under the same thread or under the same categories. This is how grouping is done.

Key Points: The email grouping system as well as email summarization system relies on a simple algorithm, but it is very complex to implement.

Dataset Used: **Enron Corpus**

Information about Dataset:

1. This is an email dataset that has approximately 500,000 emails. The dataset is publicly available for the machine learning enthusiasts
2. This dataset is widely used in e-mail related research works and publicly available

(takeaway below)

**How is this helping project??** The dataset is well known for wide variety of emails. We can use the dataset to train our model and gather the information as per our requirements within the project.

The dataset is found in Kaggle: <https://www.kaggle.com/datasets/wcukierski/enron-email-dataset>

**Methodology Used:** heuristic approach (Basic metal shortcut used to simplify a problem rather than preferring complex methods)

**Technique:** Unsupervised machine learning technique

**Working Mechanism:**

**(Function #1)**

1. As described at the beginning of the document, the model first focuses on subject and content to categorize because the crucial aspect before summarizing any email is to know the email details. Basic information like sender, receiver, subject, from date, to date, cc
2. Because these details will help us determine that the current email is thread for the previous email. The summarizer will not miss the previous information
3. Now, find most frequently used words and list of activities used.
4. Match the activities with the list of most frequently used words with their relationships (By identifying the distance and relativity) which can give a brief idea for the user about the content.
5. While assembling the sentences for summary, it selects sentences based on their occurrences in the message.

**Email Grouping System Mechanism:**

In the research paper, grouping of emails is based on user defined rules, there is an existing user’s favourite dictionary.

The Function #1 is performed which is **activity model** in this case which will determine the context of the email (whether it’s about business, work, meetings etc.)

After identifying the activity of the email, now the email is grouped under the user defined category, if now then new group is created.

(takeaway below)

**Observation:** This is one of the ways that emails are categorized when we define rules like (emails from x person to be stored under this folder)

**Performance evaluation:** Performance is measured using precision and Recall method

Recall = Total Actual Important Emails/Correctly Identified Important Emails

Precision = Total Predicted Important Emails/Correctly Identified Important Emails​

(takeaway above): We can also use the evaluation to check the accuracy of sorted mails in our project.

**Implementation that is helping in our project:**

(takeaway below)

Our project, Smart Email Mentor, aims to summarize emails over time and group them into five sections:

Business (To-Do List): Emails related to tasks, meetings, and deadlines.

Transaction & Security: Emails related to logins, security alerts, and account updates.

Order Confirmations: Emails confirming purchases, subscriptions, or renewals.

Promotional Emails: Emails containing product offers, deadlines, and coupons.

Miscellaneous: Any other emails that don’t fit into the above categories.

We can implement the below methodology from the paper

(takeaway below)

**Grouping**: We can use Email Grouping System to group under the 5 segments we have categorized.

Summarization: We can also use the summarization technique however we plan to use something more optimal model.

**Customized enhancements to the paper:**

1. Custom Activity Model: We can enhance the activity model to include specific categories like transaction & security and promotional emails.
2. Coupon Extraction: We can add a feature to extract coupons and deadlines from promotional emails.
3. To-Do List Integration: For business emails, we can extract tasks and deadlines to create a to-do list.
4. Personalization: Allow users to customize the categories and rules for grouping emails.
5. Real-Time Summarization: Implement real-time summarization as emails arrive in the inbox.
6. Integration with Calendar (in case we implement): Integrate the system with a calendar to automatically add deadlines and meetings from emails.
7. Coupon and Deadline Extraction: Enhance the summarization algorithm to extract specific information like coupons and deadlines from promotional emails.
8. We also add few more rules while developing the model

# Paper #2

DOI: <https://ieeexplore-ieee-org.rlib.pace.edu/document/9243610>

Dataset Used: **News Summary Dataset**

This dataset contains headlines, complete text, summarized text, and links to news articles. Although this dataset is primarily used for news summarization, we can adapt it for email summarization by treating email subjects as headlines and email bodies as the complete text.

**Methodology Used**

Extractive text summarization using Elmo (Embeddings from Language Models) embeddings.

**Mechanism:**

The system reads the text (in our case, email content) and splits it into sentences.

It then uses Elmo embeddings to convert each sentence into a vector that captures the meaning and context of the words.

The system compares these vectors using cosine similarity to find the most important sentences.

Finally, it selects the top 5 most important sentences and combines them to form a summary.

The system uses precision, recall, and F-measure to evaluate the quality of the summaries.

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric is used to compare the generated summary with a human-written summary (gold standard).

The system achieved higher accuracy compared to traditional methods like the TextRank algorithm.

We can implement the methodology from the paper:

(takeaway below)

Grouping:

Use Elmo embeddings to categorize emails into the five sections based on their content and subject.

For example, emails containing words like "meeting," "deadline," or "task" can be grouped under Business (To-Do List).

Summarization:

Use the extractive summarization approach to generate concise summaries of emails in each group.

For example, for Promotional Emails, the system can extract sentences containing product names, deadlines, and coupon codes.

Enhancements:

Custom Categories: We can enhance the system to include specific categories like transaction & security and promotional emails.

Coupon Extraction: Add a feature to extract coupons and deadlines from promotional emails.

To-Do List Integration: For business emails, extract tasks and deadlines to create a to-do list.

We plan to research few more techniques that will benefit our project

​