Building on the successful training run and automated evaluation, here is a comprehensive documentation section covering the **Human Evaluation** process, methodology, results, and analysis, designed for inclusion in your project report. This integrates the results from the 6_Analyze_Human_Eval.py script's output.

Human Evaluation of Generated Subject Lines

1. Rationale

While automated metrics like ROUGE provide valuable quantitative insights into the lexical overlap between model-generated subjects and human references, they have limitations, especially for abstractive tasks like subject line generation. ROUGE scores may not fully capture crucial qualitative aspects such as:

- **Semantic Relevance:** Does the subject truly reflect the core meaning or intent of the email, even if using different words?
- **Conciseness:** Is the subject appropriately brief and to the point, a key requirement for email subjects?
- **Fluency and Naturalness:** Is the subject grammatically correct and phrased in a way a human would naturally write it?
- **Informativeness:** Does the subject provide enough information for the recipient to prioritize or understand the email's context without opening it?

To address these limitations and gain a deeper understanding of the practical quality of the fine-tuned facebook/bart-large model, a human evaluation study was conducted.

2. Methodology

1. **Data Sample:** A random sample of **100 emails** was selected from the **test set** (test_cleaned_*.csv). This set was chosen as it represents unseen data for the model and contains the necessary human annotations for comparison.
2. **Subject Sources Compared:** For each sampled email body, three subject line candidates were prepared for evaluation:
    - **Model:** The subject line generated by the fine-tuned facebook/bart-large model (best checkpoint from training).
    - **Original:** The original subject line associated with the email (original_subject column).
    - **Annotation:** The first human annotation provided in the dataset (original_annotation_0 column).

3. **Raters: Two** independent human raters familiar with standard email communication practices performed the evaluation.
4. **Evaluation Criteria:** Raters evaluated each subject line based on the corresponding email body using the following criteria:
    - **Relevance/Informativeness:** How well does the subject capture the main topic or purpose of the email body? (Scale: 1=Not Relevant, 2=Somewhat Relevant, 3=Relevant, 4=Very Relevant/Specific)
    - **Conciseness:** Is the subject line appropriately short and to the point for an email? (Scale: 1=Too Long/Wordy, 2=Slightly Long, 3=Appropriate Length, 4=Very Concise)
    - **Fluency/Grammar:** Is the subject line grammatically correct, free of errors, and easy to understand? (Scale: 1=Many Errors/Unreadable, 2=Some Errors/Awkward, 3=Mostly Fluent, 4=Perfectly Fluent)
5. **Procedure:**
    - An evaluation sheet (human_evaluation_sheet_input_*.csv) was generated containing the 100 sampled email bodies. For each email body, the three subject candidates (Model, Original, Annotation) were presented in a **randomized order** (labeled A, B, C) to ensure a **blind review** (raters did not know the source of each subject).
    - Separate copies of the sheet were provided to each rater.
    - Raters independently assigned scores from 1 to 4 for each criterion (Relevance, Conciseness, Fluency) to each subject line candidate based on the provided email body.
    - Completed rating sheets (Rater1_eval.csv, Rater2_eval.csv) were collected.

3. Results

The collected ratings were analyzed using the 6_Analyze_Human_Eval.py script.

3.1. Inter-Annotator Agreement (IAA):

To assess the consistency and reliability of the human ratings, Cohen's Kappa was calculated for each criterion between the two raters. Krippendorff's Alpha was skipped as the required library was not installed.

- Cohen's Kappa (Relevance): **0.5458**
- Cohen's Kappa (Conciseness): **0.4266**
- Cohen's Kappa (Fluency): **0.4811**

According to standard interpretations (e.g., Landis & Koch), these Kappa values indicate **Moderate Agreement** between the raters across all criteria. This suggests the raters had a generally shared understanding of the criteria, making the average scores reasonably reliable, while also reflecting the inherent subjectivity in evaluating language quality.

3.2. Average Scores by Source:

The average scores (on the 1-4 scale) for each criterion, broken down by the source of the subject line, are presented below:

Source	Criterion	AvgScore	StdDev	NumRatings
Annotation	Conciseness	3.48	0.739	100
Annotation	Fluency	3.69	0.575	100
Annotation	Relevance	3.54	0.721	100
Model	**Conciseness**	**3.18**	**0.715**	**100**
Model	**Fluency**	**3.28**	**0.671**	**100**
Model	**Relevance**	**3.24**	**0.776**	**100**
Original	Conciseness	3.59	0.686	100
Original	Fluency	3.79	0.501	100
Original	Relevance	3.68	0.610	100

(Data from human_evaluation_summary_stats_*.csv)

4. Analysis & Discussion

- **Model vs. Human Performance:** The results consistently show that both the Original subjects and the human Annotation subjects received higher average ratings than the Model-generated subjects across all three criteria (Relevance, Conciseness, Fluency). The Original subjects were rated highest overall.
- **Model Quality:** Despite being rated lower than human examples, the model's average scores are all above 3.0, indicating that it generally produces subjects perceived as **better than "Fair"** and leaning towards "Good" quality on average. The lowest average score for the model was in **Relevance (3.24)** and **Conciseness (3.18)**, suggesting these might be weaker areas compared to Fluency (3.28).
- **Comparison with ROUGE:** The fine-tuned BART-large model achieved significantly higher ROUGE scores than the T5-small model on the test set (e.g., ROUGE-L 0.346 vs. 0.286). The human evaluation confirms that BART-large produces reasonably good

subjects (\u003e3.0 average scores), supporting the decision based on automated metrics, although it also quantifies the remaining gap to human-level performance.

- **Generated Length:** The automated metrics showed a median generated length of 22 tokens for BART-large on the test set. This longer length likely contributes to the lower average Conciseness score (3.18) compared to human subjects (Original: 3.59, Annotation: 3.48).
- **Qualitative Insights:** Further analysis of the specific examples saved in human_evaluation_qualitative_examples_*.csv (comparing high-scoring vs. low-scoring model outputs against human versions for the same email body) is necessary to understand the specific failure modes (e.g., genericness, missing key information) and successes of the model.

5. Conclusion from Human Evaluation

The human evaluation confirmed that the fine-tuned facebook/bart-large model is capable of generating email subject lines that are generally relevant, concise, and fluent, typically achieving ratings between "Fair" and "Good". However, it does not consistently reach the quality level of subjects written by humans, particularly regarding relevance and conciseness. The moderate inter-annotator agreement highlights the subjective nature of the task but provides confidence in the overall trend showing human superiority. The evaluation suggests that while the model is functional, future work could focus on improving relevance and exploring techniques to generate more concise outputs, potentially by adjusting generation parameters or further model tuning.

Human evaluation of sample 100 emails for generated subject lines.

script6_human_eval
uation_iaa_Raters1_

**script6_human_eval
_analysis_log_Raters**

Python Program

**5c_manual_rating_d
oc_to_csv.py**

6_analyze_human_eval_v3.py

5_Generate_Human_Eval_Prep- program

6-Analyze Human Evaluation Results -