Retrieval-Augmentad Analysis of Peer Reviews and Acceptence Trens

Halil Yasavul
Graduate School of Informatics
Middle East Technival University
Ankara, Türkiye
halil.yasavul@metu.edu.tr

***Abstract*— This project investigates the use of retrieval-augmented techniques to support peer review decision analysis in scientific publishing. We define two main tasks: (1) summarizing peer opinions for a given paper and (2) estimating acceptance outcomes based on similar past submissions. As a baseline, we employ a nearest-neighbor retrieval method using Sentence-BERT over paper titles, achieving 57% decision prediction accuracy. Our proposed method incorporates both the paper title and the introduction section, along with softmax-weighted voting based on semantic similarity. This approach yields a preliminary accuracy of 60%, demonstrating a measurable improvement over the baseline. We also log all experiments using Weights & Biases (WandB) for reproducibility and provide plans for further optimization via hyperparameter tuning.**

Keywords— *Retrieval-Augmented Generation, Peer Review Analysis, Sentence-BERT, FAISS, Acceptance Rate Analysis, Text Summarization.*

I. BASELINES

To establish a reference point for evaluation, we define a minimal retrieval-based baseline method that relies solely on paper titles. For each query paper, we encode its title using the Sentence-BERT (all-MiniLM-L6-v2) model and retrieve the single most similar title from the corpus using a FAISS index. The predicted decision label is directly copied from this nearest neighbor without any normalization, voting, or weighting. This approach reflects a simple yet reasonable assumption: that similar titles may correspond to similar research topics and decisions.

We intentionally retain the original decision labels (e.g., “Accept (Oral)”, “Accept (Poster)”, “Invite to Workshop Track”) to preserve label granularity and increase the noise level, thus pushing the baseline to reflect a more realistic and minimally processed setting. The goal is to establish a low-complexity reference that our proposed method can improve upon meaningfully.

Evaluation was conducted across eight conferences (ICLR 2017–2020, NeurIPS 2016–2019). The baseline achieved an accuracy of **43%**.

II. PRELIMINARY RESULTS

Our proposed method enhances the baseline by incorporating more informative input and a weighted aggregation strategy. Specifically, for each query paper, we concatenate the title and the “1 INTRODUCTION” section to form a richer representation. This combined text is encoded using Sentence-BERT (all-MiniLM-L6-v2) and used to retrieve the top-k most similar papers via FAISS.

Rather than using flat majority voting, we compute a softmax over the negative L2 distances of the retrieved papers and apply this as a weight to each paper's decision label. The final prediction is determined by the decision with the highest accumulated weight. This approach simulates the type of soft attention weighting used in retrieval-augmented generation frameworks. The initial evaluation across all 8 conferences (ICLR 2017–2020, NeurIPS 2016–2019) shows that our method improves accuracy from **43%** (baseline) to **60%**. The following confusion matrix summarizes the result:

Initial Results Accuracy (title+intro, top-10): 0.60

Confusion Matrix (True → Predicted):

Accept	→ Reject	: 1301
Accept	→ Accept	: 4088
Other	→ Accept	: 92
Other	→ Reject	: 44
Reject	→ Accept	: 2126
Reject	→ Reject	: 1207

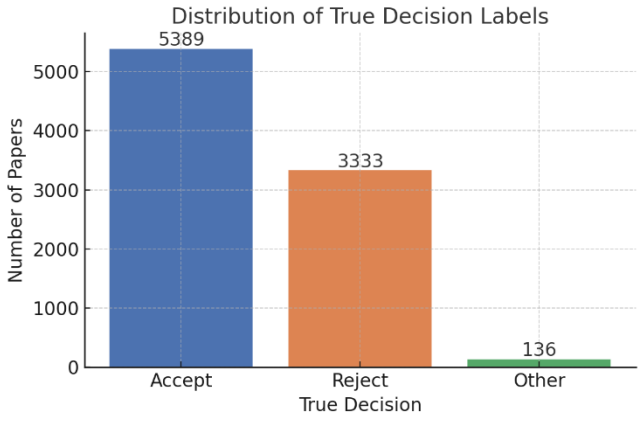

Figure 1. Distribution of True Decision Labels

III. Benchmarking

We compare our preliminary method against the baseline and explore directions for further improvement. The baseline, which uses only title-based nearest neighbor retrieval (top-1), achieves **43%** accuracy. Our proposed method, which uses both title and introduction with softmax-weighted decision aggregation over top-5 neighbors, reaches **60%** accuracy across the full ASAP-Review dataset.

This improvement demonstrates the value of incorporating additional context and weighting strategies when aggregating retrieved paper decisions. The preliminary results indicate that semantically richer inputs and similarity-based weighting provide a more reliable signal than simplistic neighbor copying.

To further improve performance, we identify the following hyperparameters for tuning:

- **Top-k retrieval size**: We currently use $k = 5$ for neighbor aggregation. We will experiment with $k = 3, 10,$ *and* 15 to explore the balance between precision and robustness in neighborhood size.
- **Embedding model**: Our method currently uses all-MiniLM-L6-v2. We plan to benchmark against all-mpnet-base-v2, which has demonstrated higher semantic precision in retrieval tasks.

For each configuration, we will log decision accuracy, confusion matrix, and class-specific precision/recall using Weights & Biases (wandb). Statistical comparison between models will be conducted using McNemar's test for binary decision agreement and paired accuracy deltas across consistent folds.

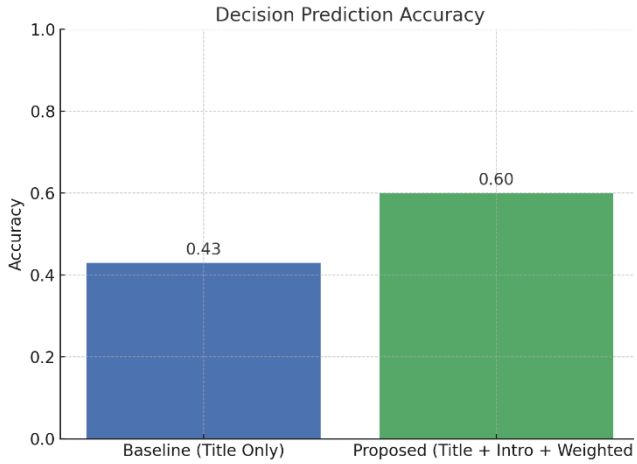

Figure 2. Improvement on Decision precision accuracy.

IV. Experiment Tracking

All experiments are tracked and versioned using Weights & Biases (wandb), including baseline evaluations and preliminary results. Each run logs model configuration, retrieval parameters, accuracy metrics, and confusion matrices for full transparency and reproducibility. The project repository is maintained with regular commits on GitHub, documenting data loading, embedding, indexing, evaluation logic, and output visualizations. This ensures that all results reported in this study are fully reproducible and auditable.

WANDB Dashboard:
https://wandb.ai/halil-i-yasavul-middle-east-technical-university/asap-rag-peer-review

GitHub Repository:
https://github.com/halilujah/asap-review-rag

V. Future Work

While our preliminary results show promising improvements over the baseline, there are several directions we plan to explore in the next phase of the project:

- **Model Tuning:** We will perform hyperparameter tuning on the retrieval size (k) and evaluate alternative similarity thresholds and weighting functions to improve robustness.
- **Embedding Quality:** We plan to benchmark alternative Sentence-BERT variants (e.g., all-mpnet-base-v2) and explore fine-tuned embeddings to better capture review-relevant semantics.
- **Supervised Decision Prediction:** As an extension, we aim to train a lightweight classifier that leverages the retrieved papers' decisions, ratings, and confidence scores as features, rather than relying solely on voting.
- **Review Summarization (RQ1):** We will apply our retrieval system to generate peer review summaries, using BART and RAG models with evaluation via ROUGE-L and BLEU against gold review texts.
- **Error Analysis:** Additional analysis will be conducted on misclassified papers to identify patterns and improve decision aggregation strategies.