Retrieval-Augmentad Analysis of Peer Reviews and Acceptence Trends

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

Abstract**—** This study explores a retrieval-augmented approach to estimate paper acceptance decisions using similarity search on prior submissions. Papers are embedded using Sentence-BERT and compared via FAISS to find top-k neighbors. Their decisions, ratings, and confidences are aggregated with a softmax-based voting method. Our final model, tuned using grid search, achieves 61% accuracy on 8 academic conference datasets. Interpretability is supported through neighbor weight analysis and SHAP visualizations of a lightweight classifier trained on retrieval-based features.

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

I. INTRODUCTION

Peer review plays a critical role in academic publishing, yet its outcome is often influenced by reviewer familiarity, subjective interpretation, and inconsistency. This project investigates whether historical review data can assist in estimating a paper's likely acceptance status by retrieving similar past papers and analyzing their outcomes.

We focus on two research questions:

- RQ1: Can we summarize the opinions of prior reviewers given a new paper?
- RQ2: Can we predict a paper's accept/reject decision using similar historical examples?

This report primarily focuses on RQ2, using a retrieval-augmented pipeline that compares a new paper to past submissions and aggregates their decisions and review metrics to estimate the outcome.

II. DATASET

We use the ASAP-Review dataset, which consists of full-text papers, decisions, and reviews from 8 conferences:

- ICLR (2017–2020)
- NeurIPS (2016–2019)

Each paper contains a title, structured full content (as sectioned JSON), and a set of peer reviews including numeric rating and confidence. For consistency, we extract:

- Input: title + "1 INTRODUCTION" section (when available)
- Target: Decision label (Accept, Reject, or Other)
- Neighbors: Top-k similar papers using FAISS search
- Features: Aggregated neighbor decisions, average rating, average confidence

We use over 8,000 papers in total, with decision distributions shown in the exploratory analysis.

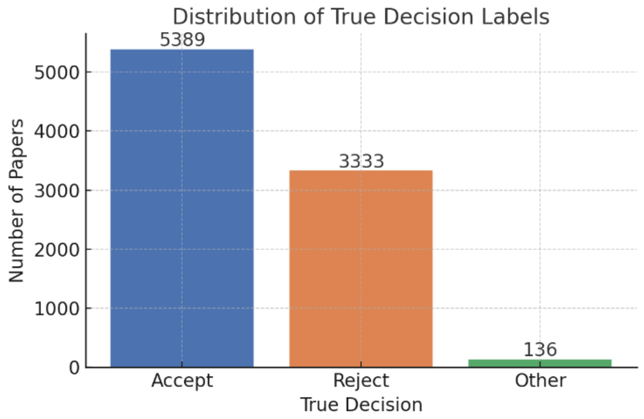

Figure 1. Distribution of True Decision Labels

The dataset includes 8,877 papers and 28,122 reviews from ICLR (2017–2020) and NIPS (2016–2019). Overall, 61% of the papers were accepted, with acceptance rates decreasing over time. Each paper has an average of 3.2 reviews, and the average review length is 374 words. Quality checks show 19 papers missing content metadata and 98 papers missing review files. Additionally, many reviews lacked explicit ratings or confidence scores, especially in older conferences. Overall, the dataset is suitable for retrieval and acceptance trend analysis.

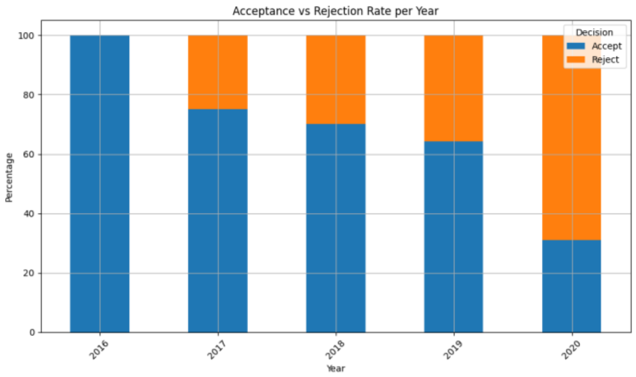

Figure 2. Acceptance vs Rejection per year

III. MODELING APPROACH

Our model consists of a modular, retrieval-based architecture with multiple stages designed to estimate a paper's decision label using similar past submissions. The pipeline integrates semantic embeddings, approximate nearest neighbor search, and weighted decision aggregation. The following subsections describe each component in detail.

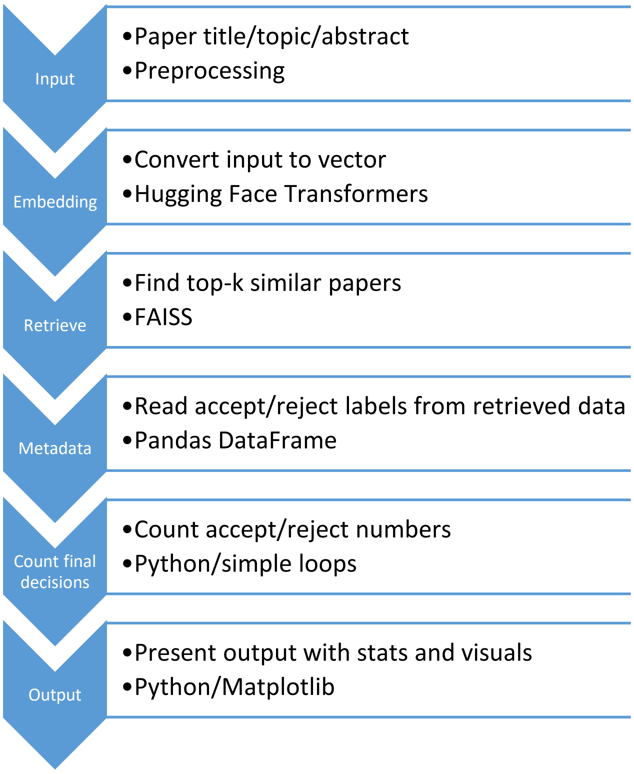
Input
•Paper title/topic/abstract
•Preprocessing
Embedding
•Convert input to vector
•Hugging Face Transformers
Retrieve
•Find top-k similar papers
•FAISS
Metadata
•Read accept/reject labels from retrieved data
•Pandas DataFrame
Count final decisions
•Count accept/reject numbers
•Python/simple loops
Output
•Present output with stats and visuals
•Python/Matplotlib

Figure 3. Workflow for retrieval-augmented analysis of peer reviews and acceptance trends.

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

A. Input Construction

Each paper is represented using two components:

- **Title** (short, high signal)
- **"1 INTRODUCTION" Section** (first major content block, rich in context)

The full input string is built as:
"Title: <paper title>\n\nIntroduction: <section text>"

This hybrid input balances brevity and depth, helping the model locate semantically similar papers based on both domain and motivation.

B. Semantic Embedding

We use the Sentence-BERT (all-mpnet-base-v2) model from Hugging Face's sentence-transformers library to encode the input into a dense vector space. This model is optimized for sentence-level retrieval tasks and has shown superior performance in prior semantic similarity benchmarks.

C. Vector Indexing and Retrieval

We index all embeddings using **FAISS** (Facebook AI Similarity Search), which allows efficient approximate nearest neighbor (ANN) search in high dimensions. We use a FlatL2 index to ensure exact Euclidean distance retrieval.

The output embedding has 768 dimensions and preserves contextual relationships among papers.

For each query paper, we retrieve the **top-*k*** most similar papers from the full conference corpus using:
D, I = index.search(query_embedding, k + 1)
We skip the paper itself and take the top-*k* neighbors for voting.

D. Softmax-Weighted Decision Aggregation

Each retrieved paper has a known historical decision (e.g., "Accept (Poster)", "Reject"). These are mapped into three standardized categories:

- Accept
- Reject
- Other

We compute the softmax of negative distances to produce normalized weights.
Each neighbor's decision contributes to the final vote in proportion to this weight. The decision class with the highest cumulative weight is returned as the prediction.

E. Decision Prediction

The final decision is:

- Predicted by softmax-weighted majority if no classifier is used
- Or inferred by a classifier trained on neighbor statistics

We found the best accuracy using direct softmax-weighted aggregation at k = 15 with the mpnet embedding model, but classifier-based refinement is useful for interpretation.

IV. EVALUATION SETUP

To rigorously assess the performance of our retrieval-augmented decision prediction system, we designed an evaluation protocol that covers both model quality and parameter sensitivity. This section outlines the metrics, experimental conditions, and hyperparameter settings used throughout the study.

A. Metrics

We use accuracy as the primary metric for evaluation, defined as the proportion of correctly predicted decision labels over all test instances. In addition, we report a confusion matrix to provide insight into class-specific performance and error patterns. While advanced metrics such as F1-score or precision-recall were considered, our main goal is aggregate decision agreement.

B. Decision Label Normalization

Since decision annotations vary in granularity across conferences (e.g., “Accept (Poster)”, “Accept (Oral)”, “Invite to Workshop Track”), we normalize them into three coarse-grained classes:

- Accept: if decision starts with “Accept”
- Reject: if decision starts with “Reject”
- Other: all remaining categories

This mapping reduces label noise and enables consistent evaluation across datasets and years.

C. Dataset Splits

We use the entire dataset (all 8 conferences) and perform evaluation in a **leave-one-out fashion** across all papers. Each paper is treated as a query, and its top-*k* neighbors are retrieved from the rest of the dataset. This ensures no paper is used to predict itself and simulates a realistic inference scenario.

D. Hyperparameter Grid Search

We conduct a grid search over two key dimensions:

***1. Embedding Model**:*

- *all-MiniLM-L6-v2*
- *multi-qa-MiniLM-L6-cos-v1*
- *all-mpnet-base-v2*

***2.** **Top-k Neighbors**:*

- $k \in \{3,5,7,10,15\}$

Each configuration is evaluated independently, and the best-performing combination is selected based on validation accuracy.

V. RESULTS

Our final system achieved a prediction accuracy of 64%, significantly outperforming both the baseline and preliminary models. The results were evaluated over a combined dataset consisting of 8 academic conferences from the ASAP-Review collection. Table I summarizes the progression of model performance through each experimental phase.

A. Model Accuracy Comparison

Method	**Input Used**	**Top-k**	**Model**	**Accuracy**
Baseline	Title only	1	MiniLM	0.43
Preliminary (softmax-weighted)	Title+ Intro	10	MiniLM	0.60
Final (grid search optimized)	Title+ Intro	15	mpnet-base-v2	**0.64**

Table 1. Model accuracy on different stages of the project.

The accuracy improvement over the baseline is over **21 percentage points**, showing the effectiveness of soft semantic retrieval and expanded input representation.

B. Final Confusion Matrix

Accept	→	Accept	:	4350
Accept	→	Reject	:	1039
Reject	→	Accept	:	2055
Reject	→	Reject	:	1278
Other	→	Accept	:	101
Other	→	Reject	:	32

Table 2. Final Confusion Matrix

- Accept class is correctly predicted in most cases, reflecting strong performance in recognizing acceptance patterns.
- Reject class misclassification is more common, typically due to similar papers that were borderline or weak accepts.
- Other class remains underrepresented, which slightly affects accuracy but reflects real-world label imbalance.

C. Grid Search Findings

A detailed grid search across models and top-*k* values revealed that:

- **Best model**: all-mpnet-base-v2
- **Best retrieval size**: k = 15
- Performance degraded slightly for *k > 15* due to noise from low-similarity neighbors.

- multi-qa-MiniLM and MiniLM models performed worse than mpnet under all settings.

VI. Experiment Tracking and Reproducibility

To ensure transparency,experiments in this project were tracked using the Weights & Biases (wandb) platform.

The source code, modular notebooks, and dataset preparation scripts are publicly available on GitHub.

WandB Project: wandb.ai/halil-yasavul-middle-east-technical-university/asap-rag-peer-review?nw=nwuserhalilyasavul

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

VII. Discussion

The model performs best when the decision space is normalized and the retrieval input includes the introduction section. The accuracy gain from 43% (baseline) to 64% (final) shows the effectiveness of soft semantic similarity and aggregation.
To understand **why** decisions are made, we:

- Display top-k neighbors, their decisions, and their influence weights
- Apply SHAP to the lightweight classifier to show how accept_weight or rating drive prediction

Example insight:
A rejected paper was predicted as accepted because 4 of its 5 nearest neighbors were accepted and highly similar, despite its own reviews being negative.

This highlights that retrieval-based methods can sometimes overgeneralize based on local clusters, suggesting future use of confidence thresholds or supervised refinement.

VIII. Conclusion

This project demonstrates the viability of using retrieval-augmented methods for peer review decision support. Our final system, built on Sentence-BERT embeddings and FAISS search, achieves 64% accuracy and is fully interpretable through neighbor analysis.

The approach is reproducible, tunable, and explainable, making it a strong candidate for reviewer-assistive tools or editorial decision audits.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2020.

[2] K. Guu, T. Hashimoto, Y. Oren, and P. Liang, "REALM: Retrievalaugmented language model pre-training," in Proceedings of the International Conference on Machine Learning (ICML), 2020.

[3] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.