KDSH

2025

TEAM: DEEPNET DYNAMOS

Table of CONTENT

OVERVIEW

AI-Driven Research Paper Evaluation and Classification

The Pathway Framework seeks to address the inefficiencies of the manual evaluation process for research papers in academic publishing by developing an AI-driven system. Manual reviews are labor-intensive and time-consuming, requiring significant expertise, and are further hindered by subjectivity, which often leads to inconsistencies in assessing the publishability of papers. As submission volumes continue to grow, the need for an automated and objective approach has become increasingly urgent.

This initiative aims to leverage advanced AI techniques to streamline the evaluation process. By automating critical tasks such as assessing the publishability of research papers and recommending the most suitable conferences for submission, the system ensures fair and consistent evaluations. The project represents a step toward improving efficiency and objectivity in academic publishing, addressing the growing challenges posed by manual review processes.

OBJECTIVE

PUBLISHABILITY ASSESSMENT

Develop a framework to classify research papers as "Publishable" or "Non-Publishable" by identifying critical issues such as flawed methodologies, incoherent arguments, or lack of evidence.

CONFERENCE SELECTION

Design a system to analyze research papers and recommend the most suitable conferences for submission, aligning with the scope, themes, and standards of various academic platforms.

ENHANCED OBJECTIVITY

Ensure consistent and fair evaluation by reducing the subjectivity involved in manual assessments through AI-driven methodologies.

STREAMLINED PROCESS

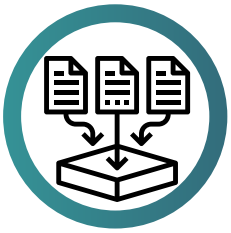
Improve the efficiency of the evaluation process, enabling faster and more reliable assessments to handle increasing submission volumes effectively.

SCALABLE FRAMEWORK

Create a solution adaptable to different academic disciplines, allowing for broader applicability across diverse research domains.

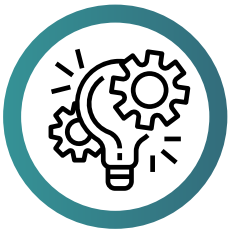
DATASET CURATION

Key steps in our data curation:



DATA PREPROCESSING

Research papers were parsed from PDF format, removing irrelevant content and ensuring clean text for analysis.



CONTENT ANNOTATION

Using a large language model, the content was automatically classified into five key sections—Introduction, Abstract, Methodology, Results, and Conclusion—ensuring structured data for training.

LABELING

Papers were labeled as "Publishable" or "Non-Publishable" based on academic standards, and suitable conferences were identified for publishable papers.

DATA AUGMENTATION AND BALANCING

To address class imbalances, data augmentation techniques were applied, ensuring a balanced and accurate dataset for improved model performance. Quality control measures were also implemented to maintain accuracy and consistency in the data.

PATHWAY INTEGRATION

This section explores how the Pathway Framework was integrated into the data curation process for research paper evaluation.

WORKFLOW AUTOMATION WITH PATHWAY

- Pathway was integrated into the workflow to automate the extraction, classification, and processing of research paper content for evaluation.
- The integration aimed to streamline manual tasks and improve the efficiency of evaluating research papers.

SEAMLESS CONNECTION TO GOOGLE DRIVE

- Pathway facilitated the connection to Google Drive, allowing for automatic retrieval of research papers stored in the drive.
- The framework also enabled the extraction of textual data from the PDF files, eliminating the need for manual parsing.

TEXT CLASSIFICATION AND ORGANIZATION

- Pathway utilized large language models to automatically categorize the extracted text into sections like Abstract, Methodology, Results, and Conclusion.
- This automated classification ensured that the papers were parsed efficiently, with minimal manual intervention.

STRUCTURED DATA FOR MODEL TRAINING

- The processed papers were used to train machine learning models for tasks like publishability prediction and conference recommendation.
- Pathway transformed raw text into structured data, making it easier to analyze and apply machine learning techniques.

EFFICIENT AND CONSISTENT EVALUATION PROCESS

- The integration enabled a faster and more consistent evaluation process by reducing the reliance on human reviewers.
- Pathway provided an automated framework that ensured uniformity in evaluations, improving the overall objectivity of the process.

FLEXIBILITY AND SCALABILITY OF THE PATHWAY FRAMEWORK

- Pathway's integration provided a flexible pipeline capable of handling changes in datasets or evaluation criteria.
- The system was easily adaptable to meet the evolving needs of academic publishing, ensuring long-term scalability.

PATHWAY INTEGRATION

WORKFLOW AUTOMATION WITH PATHWAY

↓

SEAMLESS CONNECTION TO GOOGLE DRIVE

↓

TEXT CLASSIFICATION AND ORGANIZATION

↓

STRUCTURED DATA FOR MODEL TRAINING

↓

EFFICIENT & CONSISTENT EVALUATION PROCESS

↓

FLEXIBILITY & SCALABILITY OF THE PATHWAY FRAMEWORK

PUBLISH PREDICTION

PDF PARSING

Initially, the raw research papers were parsed from PDF format, enabling the extraction of text content for further processing. This step was essential to convert the unstructured format of the papers into a usable form for analysis.

CONTENT CATEGORIZATION AND MODEL TRAINING

The extracted text was systematically categorized into five primary sections—Introduction, Abstract, Methodology, Results, and Conclusion. This categorization ensured that each paper was structured for better analysis. A labeled dataset of both publishable and non-publishable papers was then created, with labels assigned based on established academic publishing criteria. This dataset was subsequently used to train a large language model to learn and recognize patterns that differentiate publishable papers from non-publishable ones.

PUBLISHABILITY PREDICTION

Using the trained model, the system was able to predict the publishability of new research papers by analyzing the content within each of the classified sections. The model evaluated the structural and thematic alignment of the paper with academic standards, identifying potential issues that could impact its suitability for publication. This automated process aims to provide consistent and efficient publishability predictions for a large volume of research papers.

CONFERENCE SELECTION

CONFERENCE SELECTION OVERVIEW

Using the trained model, the system was able to predict the publishability of new research papers by analyzing the content within each of the classified sections. The model evaluated the structural and thematic alignment of the paper with academic standards, identifying potential issues that could impact its suitability for publication. This automated process aims to provide consistent and efficient publishability predictions for a large volume of research papers.

CATEGORIZATION OF PAPERS

The first step in the conference selection process involves classifying the papers into "Publishable" and "Non-Publishable" categories based on their content. Publishable papers are further assigned to specific conferences, including CVPR, EMNLP, KDD, NeurIPS, and TMLR, depending on their research domain. Non-publishable papers are labeled with the tag “no conference.” This classification helps in organizing the dataset and setting the stage for conference recommendation.

AUTOMATED CONFERENCE RECOMMENDATION

For the papers classified as publishable, an automated system analyzes the content to recommend the most relevant conference based on the paper’s subject matter, research focus, and technical details. The goal is to ensure that each paper is submitted to a conference that aligns with its content, increasing the likelihood of successful publication and reducing human bias in the selection process. This approach enhances the overall efficiency and objectivity of the research paper evaluation process.

CHAT-BASED REASONING

AUTOMATED CONFERENCE JUSTIFICATION WITH GEMINI API

For each publishable paper, we prompt the Gemini API to generate a 100-word justification that explains why the paper is best suited for a particular conference. The API analyzes the content, identifying key themes and research domains, and provides a coherent explanation that matches the paper to a relevant conference. This automated approach ensures objectivity and consistency, streamlining the conference selection process and minimizing human error.

CONCLUSION

our approach achieved high accuracy on the test data, on a relatively small size of the test set. Among the 135 unlabeled papers, 31 were predicted to be non-publishable, while the remaining were classified as publishable. This indicates the robustness of our system in distinguishing between publishable and non-publishable papers, despite the challenges posed by content variability. Furthermore, the conference assignment process effectively categorized the publishable papers based on their alignment with the thematic focus of different academic conferences.

We successfully implemented a novel methodology to automate the evaluation of research papers by integrating Large Language Models (LLMs) in a unique way. Our approach involved segmenting research papers into different sections, applying LLMs to each section independently, and combining their scores to make a comprehensive prediction for unlabeled papers. For conference assignment, we utilized cosine similarity between the embeddings of research papers and predefined conference embeddings to recommend the most suitable venue for each paper. This combination of techniques highlights the potential of AI-driven solutions in streamlining and enhancing the objectivity of the research paper evaluation process in academic publishing.