KHARAGPUR DATA SCIENCE HACKATHON

TEAM :
NEURAL NAVIGATORS

1. MAURYAVARDHAN SINGH (TEAM LEADER)
2. SHUBHAM KUMAR MANDAL
3. HARDIK MAHAWAR
4. AYUSHMAN PAUL

Table of Contents

1 PROBLEM STATEMENT

Task-1: Research Paper Publishability Assessment:

This task focuses on developing an automated framework to classify research papers as either **“Publishable”** or **“Non-Publishable.”** Given the increasing volume of research submissions, streamlining the peer-review process is critical to maintaining academic standards and ensuring high-quality publications. The framework utilizes **word embeddings** to convert research paper content into **vectorized representations**, allowing us to analyze and evaluate the paper’s quality effectively.

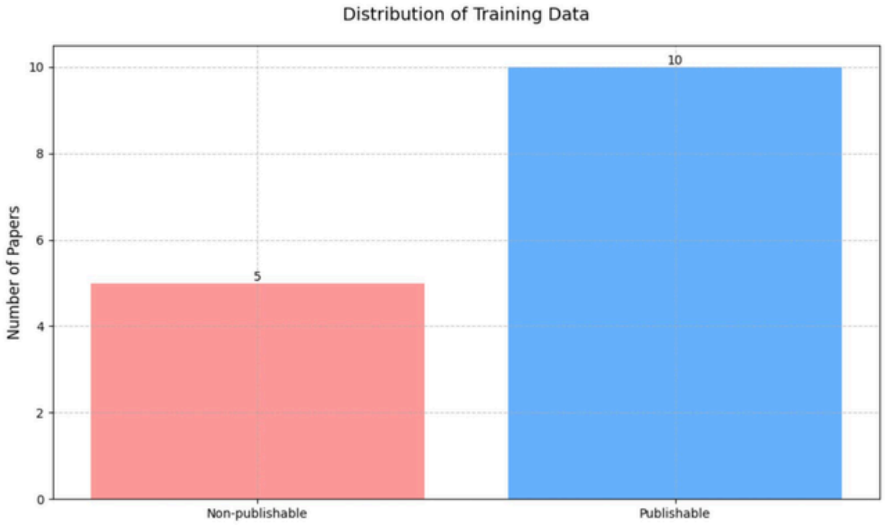
Once the paper is embedded, the system applies a **Random Forest algorithm** to classify the paper based on its adherence to key quality metrics such as coherence, methodology, and relevance. Papers are flagged as “**Non-Publishable”** if they exhibit issues like inconsistent topic alignment, lack of focus, or irrelevant content.

The objective is to develop an automated tool that helps identify papers that meet the necessary standards for publication, ensuring that only high-quality contributions make it through the publication pipeline. This system aims to improve objectivity, reduce manual effort in reviewing, and enhance the overall efficiency of the publication process across diverse academic domains.

Task-2: Conference Selection:

In this task, we aim to develop a framework that analyzes research papers and recommends the most suitable academic conferences for submission. The framework evaluates each paper’s content, research focus, and methodologies to assess its alignment with the scope and standards of prestigious conferences, including **CVPR, NeurIPS, EMNLP, TMLR, and KDD.** The system leverages **cosine similarity** to compare the paper’s content with the thematic areas covered by each conference. This enables it to match the paper to the conference that best aligns with its subject matter and research contributions. To further enhance the recommendation process, a **Large Language Model (LLM)** is utilized to generate formal justifications for the conference selection, including a **confidence score** that reflects the strength of the match.

By providing these explanations alongside the classification, the framework ensures that each paper is matched to a relevant and prestigious conference, thus facilitating a more targeted and effective submission process.



2 ANALYSIS OVERVIEW

Task-1: Research Paper Publishability Assessment:

Dataset & Purpose

- **Dataset**: **15 labeled research papers** (1 for publishable, 0 for unpublishable).
- **Objective**: Develop a model to predict whether a research paper is suitable for publication based on its content.

Data Preparation & Transformation

- **Text Extraction**: Extracted text using **PyPDF** from PDF files.
- **Preprocessing**: Cleaned data by removing unnecessary whitespace, citation numbers, and tokenized the text using **spaCy**, filtering **stopwords and punctuation.**

Feature Engineering

- **Structural Features**: Presence of sections (abstract, introduction, etc.).
- **Content Quality**: Counts of citations, equations, figures, and tables.
- **Readability**: Flesch reading ease score.
- **Technical Density:** Ratio of technical keywords.
- **Embeddings**: Used the **SPECTER** model for semantic embeddings.

Modeling & Evaluation

- **Model**: **Random Forest Classifier** with **100 estimators.**
- **Training**: **80-20** train-test split (stratified). Features were standardized using **StandardScaler.**
- **Performance**: Achieved an **F1 score of 0.800**, with a detailed classification report including precision and recall metrics.

Task-2: Conference Selection:

Dataset & Purpose

- **Dataset**: 10 labeled papers from conferences like **CVPR, EMNLP, KDD, NeurIPS,** and **TMLR.**
- **Objective**: Recommend the best-fit conference for research papers based on content similarity.

Data Preparation & Transformation

- Text Extraction: Used **LangChain's PyPDFLoader** for PDF text extraction and chunked the text using **RecursiveCharacterTextSplitter**.
- **Embeddings**: Generated using the **SPECTER** model to capture semantic meaning.
- **Domain Features**: Indicators for deep learning, computer vision, NLP, data mining, and theoretical content.

Modeling & Evaluation

- **Hybrid Approach**: Combined embedding similarity (**70% weight**) with feature-based matching (**30% weight**).
- **Justification**: Used **LLM (Ollama's llama2)** to generate academic-style justifications for recommendations.
- **Visualization**: Produced interactive charts to display similarity scores and content analysis.

Both tasks demonstrate the potential of machine learning in document analysis, with Task 1 focusing on paper publishability prediction and Task 2 on conference recommendation. The systems are designed to be modular and scalable, providing a foundation for future improvements.

3 DATA DESCRIPTION

- **Dataset Source and Scope**
  - The dataset for this project, sourced from the **Kharagpur Data Science Hackathon (KDSH)**, contains research papers and metadata from top conferences (**CVPR, NeurIPS, EMNLP, TMLR, KDD**). This diverse dataset of high-quality benchmark papers enables robust model development and evaluation for research paper classification and conference recommendation.
- **Key Variables**
  - The dataset comprises the following variables that are critical for the project:
    - Textual Content: Full-text papers in PDF format, containing sections such as the abstract, introduction, methodology, results, and conclusion.
    - Citations and References: Information about references cited in the papers, useful for assessing content quality and academic rigor.
    - Structural Information: Presence of figures, tables, equations, and other academic elements for technical evaluation.
    - Metadata:
      - Conference identifiers and labels indicating publishability and suitable conferences.
      - Research focus areas to assist in aligning papers with relevant conferences.
- **Size and Structure of the Dataset**
  - The dataset originally contains 150 research papers in PDF format.
  - Of these, 15 papers are labeled and serve as benchmarks for model training and validation.
  - The structure is hierarchical, where each paper corresponds to a unique ID, with attributes like full text, labels, and references for training and testing.
  - Data is organized chronologically to help capture trends and variations across papers and conferences.

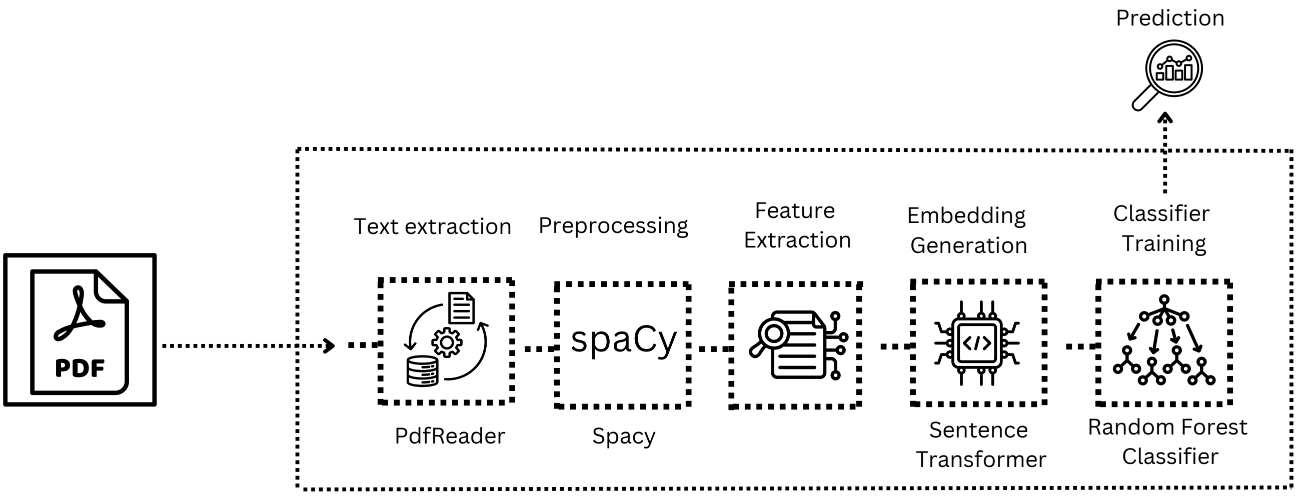
- **Data Quality and Limitations**
  - **Data Quality:**
    - The dataset underwent preprocessing to ensure clean and usable text for feature extraction.
    - Missing or incomplete fields, such as absent delay reasons or corrupted PDFs, were removed during initial cleaning.
  - **Limitations:**
    - **Sparse Labels**: Only 15 labeled papers are available for supervised learning, which limits model performance without external data augmentation.
    - **Inconsistencies**: Papers may vary significantly in length and structure, posing challenges for feature extraction and model generalization.
    - **Unlabeled Data**: While many papers are unlabeled, they require further annotation for extensive model evaluation.
    - **Domain-Specific Challenges**: Papers may contain domain-specific terminology, making semantic analysis complex without advanced NLP models.

- **Relevance of Data for Project Objectives**
  - The dataset is highly relevant to the project's goals of automating research paper evaluation and conference recommendation. Key points include:
    - **Publishability Assessment**: Features like structural completeness, citation count, and content quality help classify papers as publishable or non-publishable.
    - **Conference Matching:** Metadata and textual content align with the project’s objective of recommending conferences by matching paper themes to conference focus areas.
    - **Real-World Applications**: The dataset supports scalable analysis, enabling the development of tools that can streamline academic workflows for researchers and reviewers.

4 METHODOLOGY

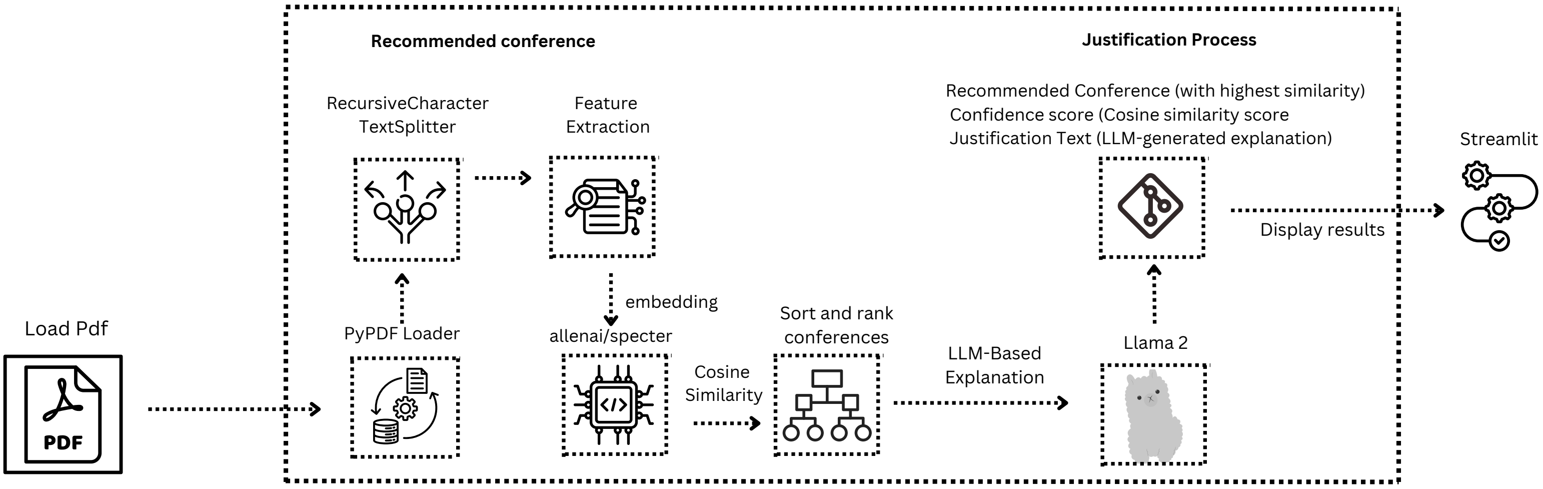
Task-1: Research Paper Publishability Assessment:

- **Approach:** Initial exploration of using LLMs and eventual adoption of Random Forest.
- **Feature Extraction**: Explanation of numerical features (e.g., abstract, citations, readability).
- **Model Training**: Overview of training the Random Forest classifier, including data splits and scaling.
- **Model Metrics:** F1 Score, accuracy, precision, recall, and detailed classification report.
- **Confusion Matrix:** Visualization of classification performance.
- **Feature Importance:** Insights into which features contributed most to predictions.



Task-2: Conference Selection:

- **Approach**: Use of cosine similarity for embedding-based conference recommendation.
- **Embedding Generation**: How SPECTER embeddings were generated for both new and reference papers.
- **Feature Matching**: Explanation of conference-specific feature extraction and similarity scoring.
- **Combined Scoring**: Description of the weighted combination of embedding and feature scores.
- **Pipeline Overview**: Flow of fetching results from Task 1 and using them for classification in Task 2.
- **Dynamic Data Handling**: Integration of CSV-based results and real-time processing.
- **Justification Generation**: Explanation of how the justification text is constructed based on features.


Recommended conference
Justification Process
RecursiveCharacter TextSplitter
Feature Extraction
Recommended Conference (with highest similarity)
Confidence score (Cosine similarity score
Justification Text (LLM-generated explanation)
Streamlit
Display results
embedding
allenai/specter
Sort and rank conferences
Load Pdf
PyPDF Loader
Cosine Similarity
LLM-Based Explanation
Llama 2
PDF

5 Data Preprocessing and EDA

Task 1: Research Paper Classification (Publishable/Not Publishable)

Data Preprocessing

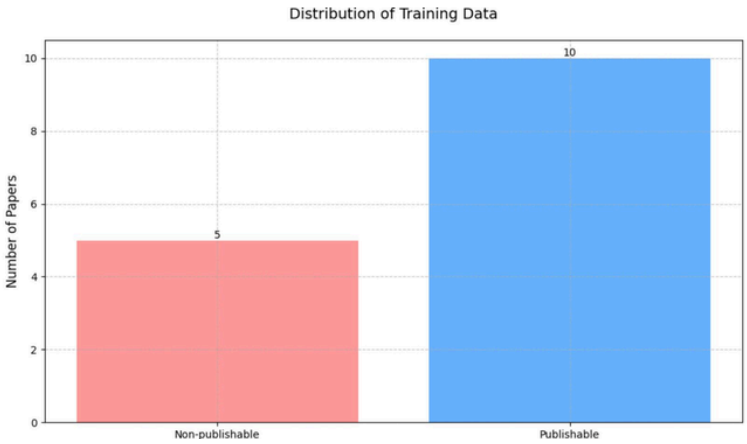
1. **PDF Text Extraction:**
   - Extracted full-text content from research papers using libraries like **PyPDF2 or PyPDFLoader**. Each page's content was concatenated into a single text string.

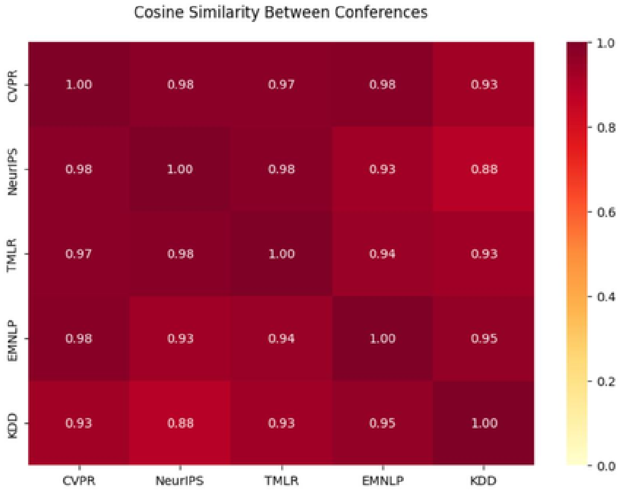
2. **Text Preprocessing:**
   - Whitespace and Citation Removal: Removed excessive spaces, newlines, and inline citations (e.g., [12] or (Author, Year)).
   - Tokenization and Cleaning: Tokenized the text using **spaCy,** removed **stop words, punctuation, and non-meaningful tokens.**
   - Feature Scaling: Numerical features such as citation counts, word counts, and readability scores were standardized using **StandardScaler.**

3. **Feature Engineering:**
   - Extracted structural and content-based features such as:
     - Section presence: Checked for sections like **Abstract, Introduction, and Conclusion.**
     - Citation density: Total number of citations per 100 words.
     - Readability metrics: Calculated the Flesch Reading Ease score.
     - Technical term frequency: Counted occurrences of keywords **(e.g., "algorithm", "methodology").**
     - Figures and tables: Detected mentions like "Figure 1" or "Table 1".

4. **Train-Test Split:**
   - Stratified sampling was used to divide the dataset into **training (80%) and testing (20%)** subsets, ensuring balanced class representation.

This image shows the distribution of papers classified as "**Publishable**" and "**Non-publishable**." The pie chart and bar graph illustrate the results of the binary classification model, which predicts the publishability of research papers.





This heatmap displays the **cosine similarity** between different conference **embeddings**. It is relevant to the task of recommending conferences for research papers based on content similarity.

6 Data Preprocessing and EDA

Task 2: Conference Selection System

Data Preprocessing

Task 2 involves two approaches:

1. **Static Recommendation Using Embeddings**
2. **Dynamic LLM-Based Justification Using llama**

For Static Recommendation:

1. **Text Extraction:**
   - Extracted the first three pages of each paper using **PyPDFLoader.**
2. **Text Preprocessing:**
   - Removed excess spaces, citations. Text was truncated to the first **512 tokens for embedding generation** using **Sentence Transformer.**

3.**Embedding Generation:**

- Generated document embeddings using the **allenai/specter model which is Sentence Transformer.**
- Conference embeddings were calculated by **averaging embeddings** of labeled papers.

4.**Feature Extraction:**

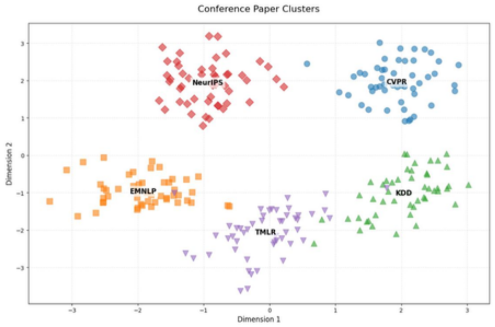
Extracted domain-specific features for each paper, including:

Keywords related to **"deep learning", "computer vision", "NLP", "data mining", and "theory".**

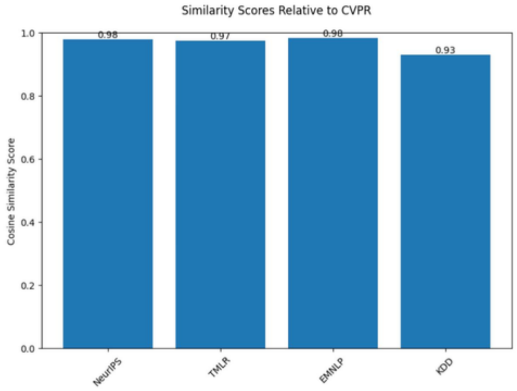
Frequency counts of domain-related terms.

5.**Similarity Computation:**

- Calculated **cosine similarity** between a new paper's embedding and the reference conference embeddings.
- Combined embedding similarity with **keyword feature overlap for final recommendations.**



This scatter plot visualizes the **clustering of papers** based on their **embeddings**, grouped by conference. It highlights how papers are distributed across different conferences, aiding in understanding the distinct research focuses of each conference.



he chart shows high similarity scores for a paper with **NeurIPS** and **EMNLP (0.98)**, suggesting strong alignment. **TMLR** is also relevant (0.97), while **KDD** is slightly lower (0.93). **NeurIPS** and **EMNLP** are top conference recommendations.

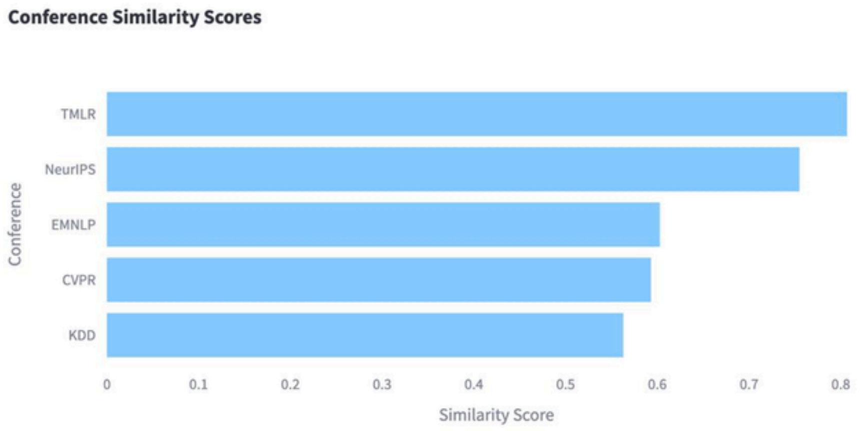
For Dynamic LLM-Based Recommendation:

1. **Prompt Construction:**
   - Used **llama LLM** with a detailed prompt to justify conference suitability based on:
     - Methodological alignment with conference focus.
     - Technical depth and experimental approach.
2. **Model Output:**
   - Generated a cohesive academic justification highlighting **why the paper aligns with the recommended conference.**

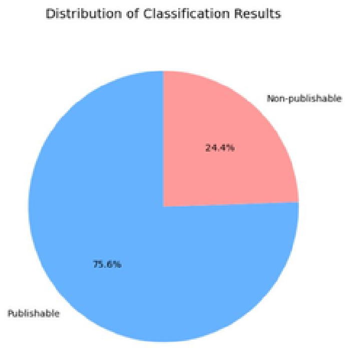
7 DATA VISUALIZATION

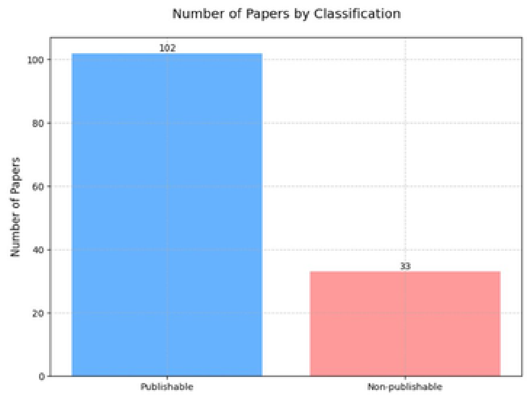
Train-Test Split: Stratified sampling was used to divide the dataset into **training (80%) and testing (20%)** subsets, ensuring balanced class representation.

- **Conference similarity score :** This bar chart displays the 5 conferences with the highest similarity scores to the uploaded research paper. The similarity score is based on both content and feature matching between the paper and the conferences. Higher similarity scores indicate a better match between the paper's research focus and the core themes of the conference.

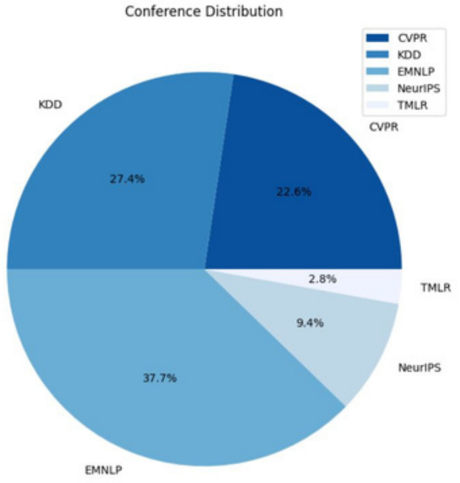


- **Distribution of Classification Results:** The distribution chart shows how papers are classified across different conferences. By visualizing the classification results, we can observe patterns such as which conference category receives the most papers and whether there is a skew towards certain research topics or themes.





- **Conference distribution:** The pie chart visually represents the distribution of research papers across various conferences. It provides an intuitive understanding of how papers are categorized based on their content, showcasing the relative proportions of each conference. This helps identify the concentration of papers in specific fields, indicating areas of high interest or specialization.



8 CHALLENGES AND LIMITATIONS

Challenges:

- **Limited Dataset Size**: One of the biggest hurdles I faced was working with a small number of labeled examples for both tasks. This made it difficult to train robust models and increased the risk of overfitting. Without enough diverse data, the models struggled to generalize to new and unseen research papers.

- **Bias in Data and Models**: With the datasets being limited and sometimes imbalanced, there's always a chance the models might become biased. This means they may not accurately represent the wide variety of research papers or conferences, which reduces their reliability.

- **Complexity in Feature Engineering**: Capturing what makes a paper publishable or suitable for a particular conference isn't easy. Some important factors, like the innovation of a paper or its real-world impact, are hard to quantify. Translating these qualitative aspects into measurable features was a challenging task.

- **High Latency:** Processing large research papers and running complex models took a lot of time. This high latency became an issue, especially for real-time applications where speed is critical.

- **LLM Hallucination:** While using language models to generate explanations and justifications, I noticed they sometimes "hallucinate," creating convincing but incorrect or irrelevant outputs. This can lead to trust issues and undermine the system's credibility.

Limitations:

- **Generalization**: With limited and specific datasets, the models often struggled to perform well beyond the examples they were trained on. This makes their real-world applications less effective.

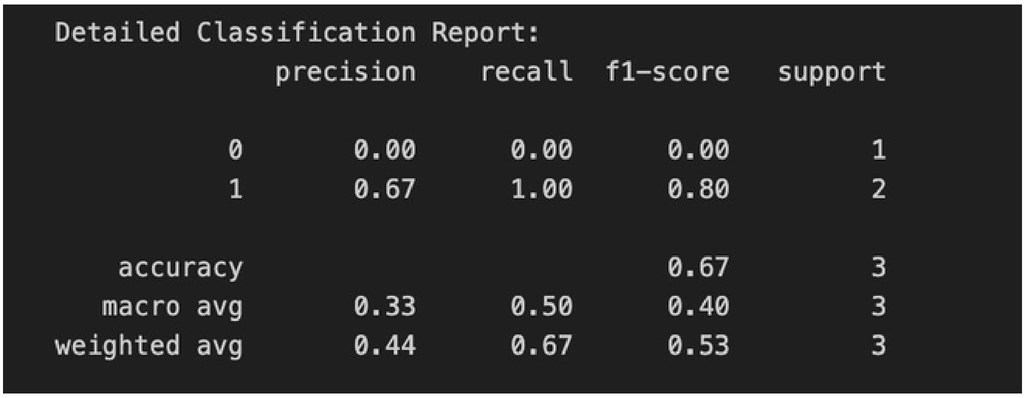
- **Interpretability**: Although the models made predictions, it wasn't always clear why they arrived at those conclusions. This lack of transparency made it harder to provide users with meaningful insights or detailed feedback.

- **Scalability**: As more papers and conferences are added, the systems may face difficulties scaling without additional computational resources or significant changes in their architecture.

- **Error Handling**: Ensuring the system could handle errors—like issues with PDF extraction or incomplete data—was challenging. Any failure in these areas could result in inaccurate predictions or system breakdowns.

The classification report provides a comprehensive evaluation of the model's performance across different conference categories. It includes:

- **Precision:** The proportion of true positive predictions among all predicted positives, indicating the accuracy of the model in classifying papers into a specific conference.
- **Recall:** The proportion of true positive predictions among all actual positives, reflecting the model's ability to capture all relevant papers for a conference.
- **F1 Score:** The harmonic mean of precision and recall, offering a balanced measure of the model's performance.
- **Support:** The number of actual occurrences of each conference in the dataset, helping assess the model's performance in terms of frequency.



```
Detailed Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         1
           1       0.67      1.00      0.80         2

    accuracy                           0.67         3
   macro avg       0.33      0.50      0.40         3
weighted avg       0.44      0.67      0.53         3
```

9 FUTURE SCOPE AND CONCLUSION

While our model provides significant insights and optimizations, several areas could enhance its effectiveness further:

1. **Enhanced Dataset Diversity:**
   - Expand the dataset by including papers from various domains (e.g., engineering, medicine, social sciences) to improve the model's generalization across disciplines.
   - Increase the number of **labelled papers** and include metadata such as reviewer comments and revisions to refine predictions.

2. **Incorporation of Advanced NLP Models:**
   - Replace or complement the **SPECTER** embedding model with more advanced models **(e.g., GPT-based embeddings, T5, or BERT fine-tuned for specific domains)** to capture deeper contextual and semantic information.

3. **Real-Time Feedback Loop:**
   - Develop a mechanism to integrate feedback from publishers and reviewers, refining the model over time with real-world outcomes.
   - Include tools to dynamically analyze new trends in research publishing, such as emerging topics or shifts in evaluation standards.

4. **Conference Recommendation Enhancements:**
   - Extend the conference recommendation system to include emerging or domain-specific conferences, ensuring broader applicability.
   - Incorporate more granular analysis of conference proceedings and align papers to tracks or themes within conferences.

5. **Scalability and Cloud Deployment:**
   - Optimize the framework for scalability by implementing parallel processing for embedding generation and feature extraction.
   - Deploy the system on cloud platforms to allow broader accessibility and real-time paper evaluations.

CONCLUSION

- The framework effectively automates research paper classification and conference recommendation, addressing the growing volume of academic outputs. While promising, limitations like small datasets and domain variability require improvement. With enhancements in NLP, scalability, and real-time learning, the system could significantly streamline publication processes, increase objectivity, and advance global knowledge dissemination.
- This system has potential to enhance objectivity and transparency in publication decisions and foster innovation by efficiently matching high-quality research with suitable platforms for dissemination.
- The initial implementation demonstrates potential, but it faces limitations such as dataset sparsity, domain variability, and challenges in handling qualitative aspects of research. Future enhancements, including larger datasets, advanced NLP techniques, and real-time learning mechanisms, can address these limitations and broaden the framework's applicability. By continually refining the model and addressing its challenges, this framework could become a robust tool for researchers, editors, and conference organizers, ultimately contributing to the advancement of global knowledge-sharing practices.

Annexure

DATASET

https://drive.google.com/drive/folders/1Z8z4craj36ighb8hzUzeM76OOgpUdsKr

RESEARCH PAPERS REFERRED -

1. https://neurips.cc/virtual/2024/poster/94429
2. https://people.scs.carleton.ca/~majidkomeili/Teaching/COMP5900-F21/TopicsPapersForPresentation.html
3. https://www.kddresearch.org/page/publications-calls#gsc.tab=0
4. https://community.nasscom.in/communities/analytics/agentic-ai-systems-opportunities-challenges-and-need-robust-governance

ONLINE FORUM DISCUSSION -

1. https://en.wikipedia.org/wiki/Conference_on_Neural_Information_Processing_Systems
2. https://www.reddit.com/r/MachineLearning/comments/bvydor/d_differences_between_ml_conferences/
3. https://docs.streamlit.io/
4. https://python.langchain.com/docs/introduction/
5. https://ollama.com
6. https://www.reddit.com/r/gradadmissions/comments/1fvev87/research_papers_not_published/