Automatic Database Construction for MLSys Papers

Aymaan Shaikh, Takuto Ban, Lijun Zhang, Hui Guan

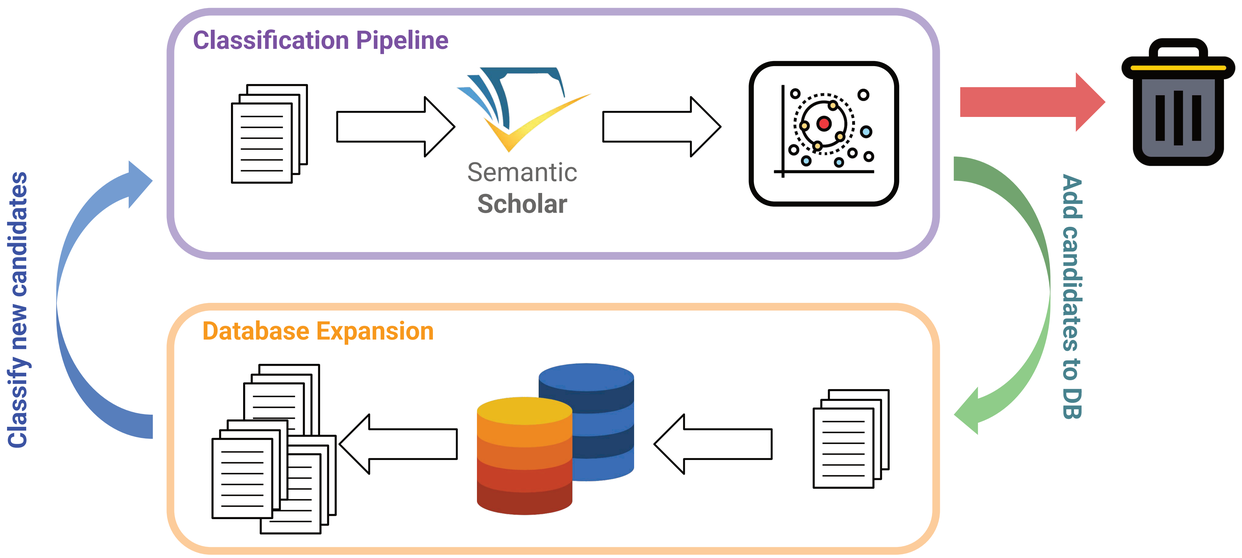
University of Massachusetts Amherst

Objectives

- **Construct a pipeline** for **curating a database** of domain-specific (MLSys) academic papers, automating as much steps as possible without human intervention
- Use machine learning techniques to **identify relevant papers** and **classify** them accordingly
- **Identify insights** about key trends, contributors, and emerging topics, to predict future research areas in MLSys

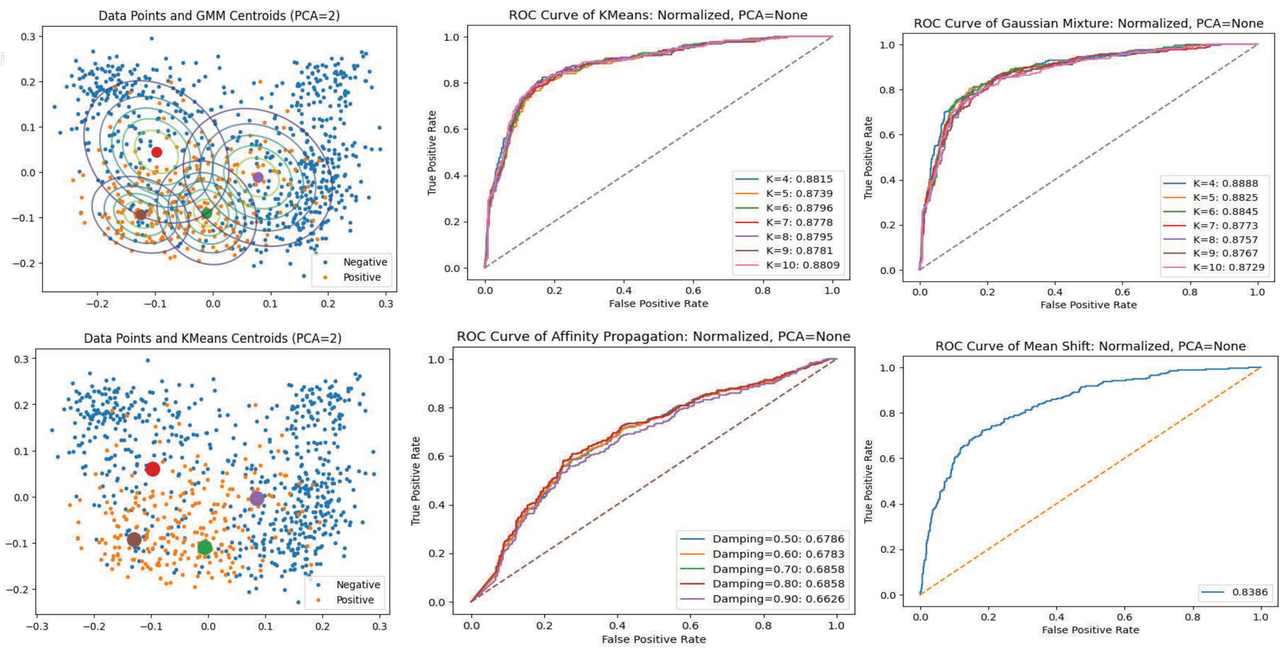
Our Approach

1. Seed Paper Collection (manual)
   a. **MLSys Conference** (2018–2024): ~450 papers
   b. **GitHub** "Awesome Systems for ML" list: ~150 papers
2. Metadata Retrieval
   a. **Semantic Scholar API** for **citation data**, **abstracts**, and **embeddings**
3. Embedding Generation
   a. **SPECTER 2**: document embedding model based on **BERT** made for academic papers
   b. **Sent2Vec**: Local, **sentence-level** embeddings aggregated to document level
4. Relevance Classification
   a. **K-Means Clustering** to identify **clusters of topics** and centroids
   b. **KNN Classifier** checks if given **embedding fall within certain threshold** to any cluster centroid
   c. If **yes** → relevant, add to DB; otherwise → irrelevant, ignore



Model Performance and Evaluation

- **K-Means** achieved **ROC-AUC of ~0.88** (unsupervised approach)
- **KNN** performed well on **curated labeled data**
- **Cosine similarity** outperformed Euclidean in KNN classification
- **Threshold tuning** improved classification precision



Trends and Insights from the MLSys Dataset

- Top research topics include **distributed training, hardware acceleration, and model optimization**
- **Topic trends** show **distributed training** steadily growing, and **hardware acceleration** surging post-2020
- **Most influential institutions**: Google, Microsoft Research, Stanford

Conclusions and Future Work

- Automate full ingestion and classification pipeline
- Extend approach to interdisciplinary domains
- Predict emerging MLSys subfields using time-series trend analysis

Selected References

- D'Cruz et al. Domain-specific Long Text Classification from Sparse Relevant Information
- Wahba et al. :Attention is Not Always What You Need: Efficient Classification of Domain-Specific Text

Contact

- Aymaan Shaikh: aishaikh@umass.edu Takuto Ban: tban@umass.edu