Publication Keyword Tagging using Machine Learning

Erol Cromwell, Maxwell Levin, Chitra Sivaraman, Sai Munikoti, Khushbu Agarwal
Pacific Northwest National Laboratory

Background

Objective

Explore natural language processing (NLP) and machine learning (ML) methods for automated keyword tagging.

Key Information

- The Atmospheric Radiation Measurement (ARM) program tracks publications using ARM data.
- Publications are manually tagged with keywords to connect them to campaigns, sites, and data products.
- These keywords help with searchability and inform ARM leadership decisions.
- **Challenges**: (1) Manual tagging is time-consuming and labor-intensive, (2) ARM processes 200+ new papers per year, with thousands of keywords available, (3) New keywords are occasionally added, requiring re-tagging of 4,500+ historical papers.

Approach

- Two approaches tested, which required no fine-tuning or pre-training, making these methods easy to implement and extend.
  - KeyBERT – Minimal keyword extraction technique using BERT (Bidirectional Encoder Representations from Transformers) text embeddings.
  - Llama 3.3 70B – A large language model (LLM) developed by Meta AI with 70 billion parameters w/ December 2023 knowledge cutoff

Datasets

Multi-Label (ML)

- **1,421 articles** from ARM Publication database
- 9 science areas, papers associated with multiple areas
- Each science area associated with several manually-derived keywords via an ARM communications expert

Multi-Class (MC)

- **69 articles** used in 2017 ARM Triennial Review
- Each paper classified into 1 of 5 science areas by domain experts

ML Science Area	# Papers
aerosol	389 (11%)
arctic	176 (5%)
boundary layer	327 (9.3%)
carbon	113 (3.2%
cloud	928 (26.3%)
model	663 (18.8%)
precipitation	185 (5.2%)
radiation	414 (11.7%)
satellite	332 (9.4%)

MC Science Area	# Papers
aerosol processes	9 (13%)
boundary layer	29 (42%)
deep convective	14 (20.3%)
mixed-phase and arctic clouds	13 (18.8%)
modeling	4 (5.8%)

Table 1: Articles per science area for multi-label and multi-class datasets

Methods

KeyBERT

- Create embeddings (numerical representation of text) for keywords and papers
- Calculate cosine similarity score between keywords and papers
- Assign paper to a science area if keyword's similarity exceeds threshold
- Threshold for science area optimized by maximizing accuracy for the science area

Llama 3.3 70b

- Two model prompt types:
  - Zero Shot (ZS): Ask model to classify paper to most relevant science area(s)
  - In-Context Learning (ICL): Provide input/output example for task (multi-class only)
- Experiment with:
  - Paper text length: Abstract (Ab), first 10K characters (10K), full text (FT)
  - Increasing size of input context window: default size (2K) and 32K tokens

Results

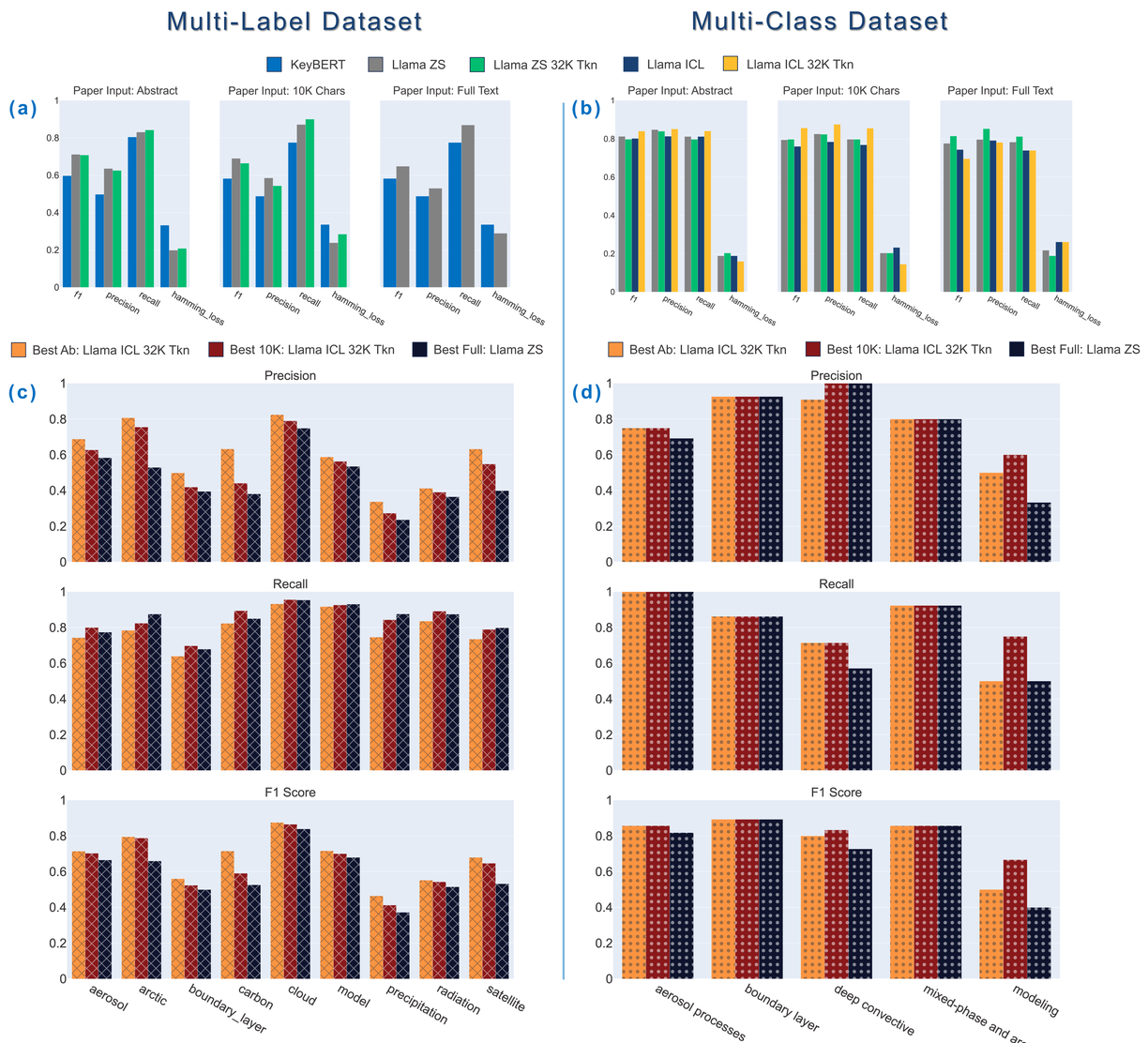

Fig 2: KeyBert and Llama 3.3 results on multi-label and multi-class dataset. Figures (a) and (b) show the overall performance of each model by paper text length. Figures (c) and (d) show performance of best model of each input type.

Key Takeaways

- **Llama outperforms** KeyBERT in **multi-label dataset**
- Llama performs worse on multi-label set compared to multi-class set
- **Increasing input context length** from default (2K) to 32K **improves performance for 10K prompts** in **multi-class set**
- ICL prompts outperform ZS prompts for abstract and 10K text inputs
- **Full text prompts struggle overall** due to paper length exceeding context window length

Next Steps

- Run prompt with different LLMs for comparison, such as DeepSeek, Mixtral, and GPT-4o
- Use in-context learning on multi-label dataset
- Increase input context window for full text prompt
- Have domain scientists validate multi-label dataset labels

Acknowledgments

ARM is sponsored by the U.S. Department of Energy's Office of Science under the Biological and Environmental Research (BER) program.

This research was performed using PNNL Research Computing at Pacific Northwest National Laboratory.

Contact Information

- Erol Cromwell, erol.cromwell@pnnl.gov
- Maxwell Levin, maxwell.levin@pnnl.gov