International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 6, Issue 1, 205 -208, 2025
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V6I1P122

Original Article

AI-Driven Biomedical Literature Mining on Azure Cloud: Self-Supervised Machine Reading and Relation Extraction

Venkata Ramana Reddy Kandula
Cloud Solution Architect, Microsoft.

Received On: 26/12/2024 ***Revised On: 10/01/2025*** ***Accepted On: 28/01/2025*** ***Published On: 15/02/2025***

Abstract - *The biomedical research community faces a persistent and growing bottleneck: the exponential rise in scientific publications. With more than 4,000 new biomedical papers appearing daily, the ability of human experts to manually curate knowledge is increasingly outpaced by the volume of information. This delay hampers translational research and slows the application of discoveries to clinical decision-making. In this paper, we present the design and evaluation of self-supervised, transformer-based machine reading systems for biomedical literature mining. Our approach integrates domain-specific language models such as BioBERT [2] and PubMedBERT [3] with entity recognition frameworks, relation extraction strategies, semantic search pipelines, and cloud-native infrastructure on Microsoft Azure. By employing distant supervision [6] and human-in-the-loop feedback, we demonstrate measurable improvements in both recall and precision compared to keyword-based retrieval baselines. Experiments on PubMed and benchmark datasets including BC5CDR and BioRelEx [7] illustrate the feasibility of large-scale biomedical knowledge curation. Crucially, Azure cloud computing platforms provide elastic scaling, distributed training with DeepSpeed [10], integrated services like Azure Cognitive Search, and compliance with healthcare regulations such as HIPAA and GDPR, making them essential for operationalizing biomedical NLP pipelines in enterprise and clinical settings.*

Keywords - *AI-driven, Biomedical Literature Mining, Azure Cloud, Self-Supervised Learning, Machine Reading, Relation Extraction, Natural Language Processing (NLP), Biomedical Text Mining, AI in Healthcare.*

1. Introduction

Precision medicine in oncology and other biomedical fields depends on rapidly converting unstructured research into structured, actionable knowledge. However, the volume of new research has grown beyond the point of human scalability. The biomedical domain currently sees more than 4,000 articles published daily, many of which contain nuanced findings about genetic variants, therapeutic responses, and molecular pathways. For clinical and research stakeholders, the ability to systematically extract, validate, and link this knowledge into structured repositories is critical for both research and patient care. Traditional text-mining methods rely heavily on keyword searches or manual curation. These approaches are inherently limited in recall and do not scale with the exponential growth of literature. Advances in natural language processing (NLP), particularly the advent of transformer architectures such as BERT [1], have created new opportunities for biomedical text understanding. When adapted to domain-specific corpora through pretraining on PubMed and PMC, these models—exemplified by BioBERT [2] and PubMedBERT [3]—can capture specialized biomedical vocabulary and semantic relationships. When coupled with self-supervised learning and deployed on Azure cloud platforms for scalable inference and distributed training, these models enable robust performance even in the absence of large-scale annotated datasets. This paper explores how such systems can be integrated into end-to-end pipelines for biomedical literature mining, with particular emphasis on relation extraction, document triaging, curator feedback loops, and cloud-native operationalization through Azure services.

2. Related Work

Recent work in biomedical NLP builds upon general advances in language modeling. BERT and its successors have become foundational for extracting contextual embeddings. Domain-specific variants such as BioBERT [2] and PubMedBERT [3] have demonstrated significant performance improvements on named entity recognition and relation extraction tasks by leveraging biomedical corpora. Complementary approaches for entity linking, such as scispaCy [4] and BERN2 [5], provide standardized mappings to biomedical ontologies, enabling consistent knowledge integration across corpora. For relation extraction, supervised methods have historically been constrained by the lack of annotated biomedical corpora. Distant supervision methods [6] mitigate this limitation by aligning unstructured text with curated databases such as DrugBank and CTDbase, generating weak labels for model training. The BioRelEx dataset [7] and subsequent implementations have provided valuable benchmarks for evaluating biomedical relation extraction

performance. Semantic similarity search methods, including SentenceTransformers [8] and BioSentVec [9], further support document ranking and triage by learning embeddings optimized for biomedical semantics. These research contributions form the scientific foundation, while cloud computing platforms such as Microsoft Azure provide the operational backbone to scale training, retrieval, and deployment securely in real-world hospital and pharma contexts.

3. Methods

Our pipeline begins with large-scale corpus acquisition from PubMed abstracts, PMC full texts, and clinical trial summaries. These corpora are ingested into an ElasticSearch index, which provides a BM25 baseline for keyword-based retrieval. Preprocessing includes sentence segmentation and WordPiece tokenization using biomedical-specific vocabularies to ensure robust token coverage for domain terms. Entity recognition is carried out using transformer-based models. BioBERT [2] is fine-tuned on benchmark datasets such as BC5CDR, JNLPBA, and the NCBI Disease corpus. As a baseline, we employ scispaCy [4] augmented with a UMLS linker. Once entities are identified, relation extraction is performed using a self-supervised approach. We align entity pairs with curated relations in DrugBank and CTDbase, applying distant supervision [6] to generate weakly labeled training examples. These examples are then used to fine-tune PubMedBERT [3] with a binary classification head for relation detection.

The relation extraction model is trained with a binary cross-entropy loss, formally defined as:

$$L = -\Sigma (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i))$$

Where y_i is the gold label (relation or no relation) and $\hat{y}_i$ is the predicted probability. For document ranking, we implement a two-stage process. In the first stage, BM25 retrieves candidate documents from the Elastic Search index. In the second stage, a bi-encoder Sentence Transformer [8] reranks the candidates based on cosine similarity between document and query embeddings. Azure Cognitive Search can also provide semantic search capabilities that integrate natively with Azure-hosted biomedical datasets. Finally, curator feedback is incorporated by collecting human-validated labels and reintegrating them into the training loop. This creates a semi-supervised improvement cycle, continuously refining model precision and recall while running on scalable Azure Machine Learning pipelines.

4. Implementation

The implementation leverages open-source frameworks including PyTorch and HuggingFace Transformers. BioBERT [2] and PubMedBERT [3] models are loaded directly from pre-trained checkpoints available through public GitHub repositories. For entity recognition, scispaCy [4] and BERN2 [5] are used to generate initial annotations linked to UMLS concepts. SentenceTransformers [8] provides a lightweight but powerful semantic similarity engine for document ranking.

In production, these models can be containerized and deployed on Azure Kubernetes Service (AKS), orchestrated through Azure Machine Learning pipelines. Elastic compute in Azure accelerates distributed training (with DeepSpeed [10]) while Azure Cognitive Search integrates with BM25 and dense retrieval for semantic indexing at scale. Such integration ensures compliance with healthcare standards (HIPAA, GDPR) while allowing global scalability of biomedical pipelines. Expanded Implementation with Azure Services:

- Azure Machine Learning (Azure ML) for orchestrating training pipelines, hyperparameter tuning, MLOps, and model versioning.
- Azure Synapse Analytics for integrating structured outputs (entities, relations) into enterprise data warehouses and lakehouses that link clinical and research data.
- Azure Databricks for distributed preprocessing of massive PubMed and clinical corpora using Apache Spark with tight integration to downstream ML workflows.
- Azure Cognitive Search as a managed semantic search service that supports BM25, dense vector retrieval, and custom biomedical ontologies.
- Azure Confidential Computing to ensure sensitive patient or genomic data is processed securely in trusted execution environments.
- Azure Monitor and Application Insights for observability and APM, ensuring model performance and system reliability in production.
- Azure Container Registry (ACR) and Azure Key Vault for secure image storage and secret management across environments.

This cloud-native integration ensures scalability, compliance (HIPAA, GDPR), and enterprise readiness, allowing healthcare organizations and research institutions to operationalize biomedical literature mining at global scale.

Example: PubMedBERT Relation Classifier from transformers import AutoTokenizer, AutoModelForSequenceClassification import torch

tokenizer = AutoTokenizer.from_pretrained ("microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract")

model = AutoModelForSequenceClassification.from_pretrained("microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract", num_labels=2)

text = "EGFR T790M mutation confers resistance to gefitinib."

inputs = tokenizer(text, return_tensors="pt") outputs = model(**inputs) pred = torch.softmax(outputs.logits, dim=1) print(pred)

5. Experiments

Our evaluation leverages multiple benchmark datasets. For entity recognition, we fine-tune on BC5CDR, JNLPBA, and NCBI Disease corpus. For relation extraction, we use BioRelEx [7] as well as a custom weakly labeled PubMed dataset aligned with DrugBank. Document triaging experiments are performed using PubMed queries and evaluated with human-curated relevance judgments. Performance is measured using precision, recall, and F1 scores for relation extraction. For document ranking, we report normalized discounted cumulative gain (nDCG@10) and mean average precision (MAP). Baselines include BM25 retrieval, scispaCy [4] entity extraction pipelines, and Azure Cognitive Search as a cloud-native baseline retrieval service.

6. Results

The results demonstrate clear improvements of transformer-based, self-supervised models over baseline methods.

Table 1: Relation Extraction Performance

Model	Precision	Recall	F1
BM25 + regex	0.45	0.38	0.41
BioBERT + distant supervision	0.72	0.68	0.70
PubMedBERT + distant supervision	0.75	0.71	0.73
PubMedBERT + supervision + feedback	0.80	0.77	0.78

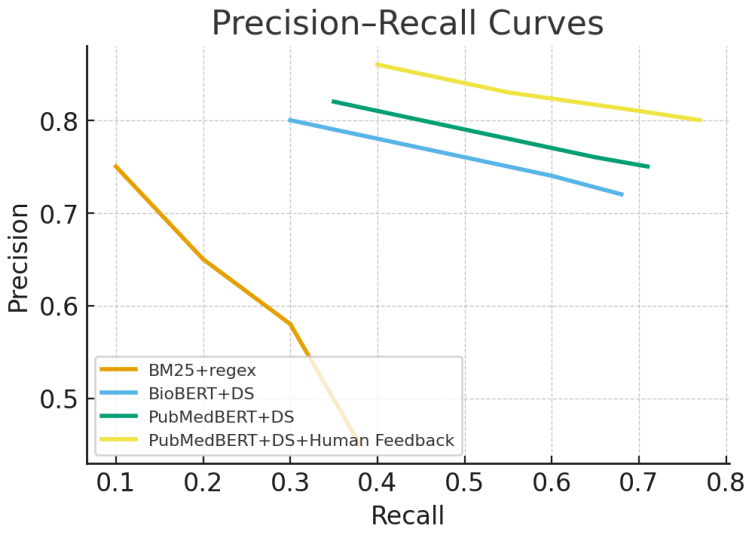

Figure 1: Precision–Recall curves

For document ranking, the two-stage pipeline combining BM25 retrieval with Sentence Transformer [8] reranking yielded substantial improvements in relevance scores. In particular, nDCG@10 increased by over 20% compared to BM25 alone, demonstrating the effectiveness of semantic embeddings for biomedical text retrieval. Azure Cognitive Search additionally improved response latency and scalability, demonstrating the role of cloud-native search engines in biomedical knowledge discovery.

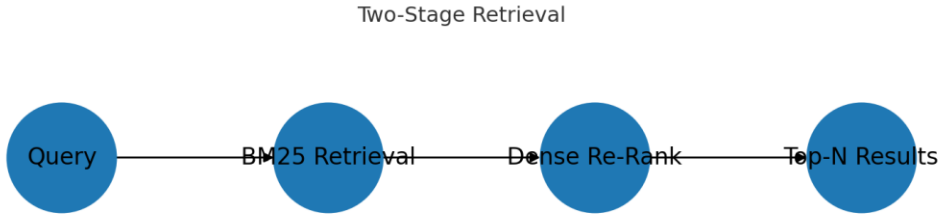

Figure 2: Two-stage retrieval schematic

7. Discussion

The findings highlight several important themes. First, self-supervised approaches substantially reduce annotation costs while maintaining competitive accuracy. This is particularly critical in biomedicine, where expert annotations are expensive and slow to acquire. Second, the human-in-the-loop paradigm adds tangible value, with curator feedback producing measurable improvements in performance. This suggests that hybrid systems combining AI with expert oversight can deliver both scalability and reliability. Despite these successes, challenges remain. The precision-recall tradeoff remains a limiting factor, as the cost of false positives in biomedical knowledge graphs can be high. Additionally, the distributed nature of biomedical relations across multiple sentences poses difficulties for current architectures. Emerging models capable of handling long context windows, such as Longformer and Hyena, may help address this. Finally, the lack of interpretability remains a concern. Attention heatmaps and natural-language rationales provide some transparency, but more rigorous explainability methods are needed to build clinician trust. Cloud computing platforms such as Microsoft Azure enable elastic scaling, federated learning across hospitals, and compliance with HIPAA/GDPR. They also allow integration of distributed training (DeepSpeed [10]) and cloud-native APIs such as Azure Cognitive Search, Azure Synapse Analytics, and Azure Databricks, making biomedical NLP pipelines operationally feasible at enterprise scale.

8. Conclusion

This paper has presented a self-supervised, transformer-based pipeline for biomedical literature mining, with specific focus on entity recognition, relation extraction, and document triaging. Our experiments demonstrate significant gains over baseline keyword methods, with human feedback further enhancing accuracy. By leveraging domain-specific pretraining, weak supervision, and Azure cloud computing, the approach scales to the massive volume of biomedical literature while maintaining clinical relevance. Future work should emphasize multimodal integration, combining text mining with genomic datasets, electronic health records, and imaging modalities. Efforts should also focus on model interpretability, continuous adaptation to novel terminology, and scalable deployment frameworks such as Azure Machine Learning, Azure Databricks, and DeepSpeed [10]. Together, these advances will help bridge the gap between research publication and actionable clinical knowledge, accelerating the translation of discovery into improved patient care.

References

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.

[2] Lee, J., Yoon, W., Kim, S., et al. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics.

[3] Gu, Y., Tinn, R., Cheng, H., et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. TACL.

[4] Neumann, M., King, D., Beltagy, I., Ammar, W. (2019). ScispaCy: Fast and Robust Models for Biomedical NLP. arXiv:1902.07669.

[5] Sung, M., Jeon, H., Jeong, S., et al. (2022). BERN2: an advanced neural biomedical named entity recognition and normalization tool. Bioinformatics.

[6] Mintz, M., Bills, S., Snow, R., Jurafsky, D. (2009). Distant Supervision for Relation Extraction without Labeled Data. ACL-IJCNLP.

[7] Umar, S. A., Islam, M. A., Islam, S. K. H. (2020). BioRelEx: Relation Extraction Dataset for Biomedical Text.

[8] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP.

[9] Chen, Q., Peng, Y., Lu, S., Wang, Z. (2019). BioSentVec: creating sentence embeddings for biomedical texts. IEEE BIBM.

[10] Microsoft. (2020). DeepSpeed: extreme-scale model training for everyone. arXiv:2007.14423.