Project Report

On

Title

Text Mining in Rheumatology Research Publication

Submitted by : Archana Senthil

UCID : 30265251

Supervised by : Dr. Mina Aminghafari

Work term : May 2025 - December 2025

Introduction

Clinical trials are the cornerstone of medical innovation and progress. They provide a structured and rigorous approach to evaluating the safety, efficacy, and broader impact of treatments and interventions. Within the realm of medical research, these trials form the foundation upon which clinical guidelines and evidence-based care are built. In the context of rheumatology, and more specifically rheumatoid arthritis, clinical trials serve as the primary mechanism through which new drugs, therapies, and treatment strategies are tested and validated. Tracking and understanding how these trials are used, cited, and interpreted within scientific literature is essential for ensuring that findings are not only accessible but also integrated effectively into ongoing research and clinical practice.

However, with the rapid growth of biomedical publications, it has become increasingly difficult to manually monitor and synthesize clinical trial information from the vast and ever-expanding corpus of scientific papers. Researchers and clinicians often face challenges in identifying which trials are being cited, how they are being discussed, and what conclusions are being drawn from them. This bottleneck limits the ability to conduct systematic reviews, meta-analyses, and evidence synthesis, all of which are critical for making informed decisions in clinical and policy settings.

This project aims to bridge this gap by developing an automated pipeline using Natural Language Processing, or NLP. NLP is a subfield of artificial intelligence that enables machines to understand and process human language. By leveraging NLP, we aim to create a system that can efficiently scan research papers, extract relevant clinical trial information, and classify the papers based on predefined attributes. Specifically, the tool will extract clinical trial identifiers such as NCT numbers, determine the first mention of each trial across the literature, and group papers based on the type of study, such as randomized controlled trials, meta-analyses, reviews, or observational studies.

The end goal is to provide researchers with a more organized and searchable database of publications tied to clinical trials. By doing so, the system will significantly reduce the manual effort involved in literature reviews, facilitate better tracking of how individual trials are interpreted across studies, and enhance the overall visibility of clinical trial usage in research. For example, identifying the first appearance of a specific trial in literature can help measure the trial's influence and reveal how its findings propagate through the scientific community.

Beyond the technical value, this project also has the potential to promote greater transparency and reproducibility in research. By offering a structured way to identify and classify trial-related documents, it ensures that critical data points are not missed and that the body of evidence supporting treatment decisions is both comprehensive and traceable.

In summary, this project applies advanced NLP techniques to address a growing need in clinical research: the ability to efficiently extract, interpret, and organize clinical trial references from

biomedical literature. In doing so, it supports the broader goal of making medical evidence more accessible, traceable, and actionable.

Project Goals

The objective of this project is to develop a robust and scalable natural language processing (NLP) pipeline to support the Rheumatoid Arthritis research team. The key goals are:

- **Text Extraction:** Automatically extract text from research papers in various formats (e.g., PDF, JSONL).
- **Clinical Trial Identification:** Detect and extract clinical trial registration numbers mentioned within the text.
- **First Use Mapping:** Identify the earliest paper and corresponding publication year in which each clinical trial was referenced.
- **Paper Classification:** Group and classify research papers based on the clinical trial referenced.
- **Study Labeling:** Label each paper with relevant metadata based on the nature of the study (e.g., observational, interventional, randomized controlled trial, etc.).

This tool aims to streamline literature analysis, enabling more efficient tracking and categorization of clinical research in the context of rheumatoid arthritis.

Roles and responsibilities

As a Research Assistant with Prof. Mina working with the Rheumatoid Arthritis team, my primary duty was to develop code aligned with the project scope and deliver a reliable, functional pipeline that supports the identification and classification of clinical trial information from research papers

My responsibilities included not only designing and implementing the codebase but also ensuring that the solution was efficient, scalable, and accurate. A critical part of this involved collaborating directly with the client to understand their needs and expectations, and to clarify requirements whenever uncertainties or questions arose. This helped ensure that the technical solution was aligned with the goals and practical constraints of the research team.

In addition to my individual technical contributions, I played an active role in a collaborative team environment. This involved participating in regular check-ins, sharing updates, helping troubleshoot technical challenges, and contributing to group discussions aimed at refining the

project's direction. By working closely with both researchers and fellow developers, I focused on enhancing team coordination and ensuring the overall quality and impact of our outcomes.

Throughout the internship, I encountered several challenges, such as inconsistencies in paper formats, ambiguities in text referencing clinical trials, and handling large-scale textual data. Part of my role was to proactively identify these challenges and either mitigate them through code or propose alternative approaches that would still meet the client's needs.

By translating project goals into actionable data workflows and providing clarity on what was feasible through automation, my contributions helped advance the overall progress of the team's data extraction and analysis capabilities.

Outcomes

I had the opportunity to work on a technically rich and interdisciplinary project focused on extracting and classifying clinical trial information from research papers relevant to rheumatoid arthritis. This experience provided me with a deeper understanding of applied natural language processing (NLP) techniques in a real-world research context and allowed me to expand both my technical and professional skill set.

Throughout the internship, I significantly improved my technical fluency in working with JSONL datasets, Python-based text processing libraries, and tools for parsing, cleaning, and extracting information from scientific literature.

Professionally, I developed stronger communication and stakeholder engagement skills. By interacting directly with domain experts and my supervisor, I learned how to translate vague requirements into actionable technical steps. Additionally, through regular team check-ins, I strengthened my ability to collaborate across disciplines, adjust to evolving priorities, and contribute meaningfully to group discussions.

A major accomplishment so far has been successfully extracting the publication date and identifying the clinical trial numbers mentioned in research papers. This involved developing a flexible extraction strategy that could accommodate inconsistent formatting and partial matches between data sources. I discovered early on that direct alignment between the database (containing known clinical trials) and the original papers was not always possible. In many cases, the trial number was either missing, formatted differently, or ambiguous. As a result, I had to come up with multiple approaches including adjusting pattern matching logic, pre-processing the text differently, and fine-tuning keyword filters to improve the detection accuracy.

Other goals such as determining when a trial was first introduced in the literature and classifying papers based on trial usage or study type are currently in progress. These tasks require deeper analysis and integration of more advanced classification logic, which I plan to build upon the initial extraction work.

On the non-technical side, I am proud of the role I played in fostering a collaborative team culture, helping peers with technical issues, and keeping an open channel of communication with both technical and non-technical collaborators.

At the start of the internship, the main objectives were:

- Build a pipeline to extract clinical trial metadata from research papers.
- Classify and group papers based on trial usage.
- Contribute to the research team's broader goal of automating evidence synthesis.

While I have made strong progress toward the first goal, the latter two are still evolving. Due to discrepancies between the client's database and real-world document structures, I had to spend more time refining the extraction phase before confidently moving on to classification. This adjustment reflects the real-world need to adapt goals in response to technical limitations and user feedback.

During the internship, I was exposed to a wide range of technical tools and methodologies. I worked extensively with Python for data extraction, cleaning, and processing. A major component of the work involved web scraping to gather structured and unstructured data from online sources. Additionally, I became comfortable with making detailed API calls, which was instrumental in retrieving and integrating clinical trial information from external databases. I also worked with regular expressions (regex) for pattern matching, Pandas for data manipulation, and Natural Language Processing (NLP) techniques to process research papers. Toward the later stages of the project, I began exploring contrastive learning approaches with LLMs (Large Language Models) for potential use in document similarity and classification tasks.

This experience enhanced my problem-solving skills and deepened my understanding of real-world AI applications. I improved my ability to debug code efficiently, manage messy data, and handle inconsistencies between sources. These are essential skills that will support my success in academic courses, especially those related to data science, machine learning, and capstone research projects. I also developed better collaboration and communication habits, which will help in group-based academic work.

One of the biggest lessons was understanding the complexity and unpredictability of real-world data projects. I learned that a solution that seems straightforward in theory often requires iterative refinement and creativity in practice. It was a strong reminder that project timelines can shift and that being adaptable is as important as being technically skilled.

I was surprised by how long and complex the work became. Initially, I believed the task of extracting and linking clinical trials would be straightforward and could be completed within three months. However, I quickly realized that inconsistencies in data formats, unexpected edge cases, and evolving client requirements made the process far more involved. The project required

multiple pivots and ongoing testing, which made the learning experience richer and more challenging.

Persistence, problem-solving ability, and a willingness to learn independently helped me overcome many obstacles during the internship. I was also highly organized, which made it easier to document changes, manage multiple files, and track progress.

Based on my self-assessment, one skill that could have been better utilized is proactive stakeholder communication. While I met client expectations, I believe more frequent updates and visual progress summaries could have further improved collaboration and faster feedback cycles.

My role directly contributed to the success of the research team by laying the groundwork for an automated and scalable system to process clinical research papers. This not only reduces manual effort but also ensures higher accuracy and consistency in trial tracking. The methods and tools I implemented can serve as a prototype for future expansion and integration into broader research workflows.

Reflection on Academic Links to Work

During my work term as a Research Assistant at the University of Calgary, I applied a wide range of technical skills and concepts that I had developed through my academic coursework. My foundational knowledge in Machine Learning and Natural Language Processing (NLP) was essential, as the project involved extracting and analyzing clinical trial identifiers from academic publications using unstructured text data. These tasks required a deep understanding of text preprocessing, regex-based pattern recognition, and working with tokenization, which I had been introduced to in earlier courses.

The theoretical understanding of language models, tokenization, and pattern recognition from class was immediately applicable to real world tasks like writing functions to extract structured identifiers from raw text. I also used knowledge from data preprocessing and cleaning to handle inconsistencies in formats, missing data, and mismatches between external databases and publication metadata.

Additionally, my academic training in programming and model development, especially the concepts taught in foundational programming and data structures classes, proved incredibly helpful. Understanding how to break down problems, write modular code using functions, and debug efficiently made it easier to implement solutions and iterate quickly. The classwork that involved working with multiple data formats such as JSON, CSV, and TXT also helped immensely, as the project required handling and parsing a variety of files. These skills would have been difficult to pick up on the fly.

One key issue I encountered was that the clinical trial data from publications did not always match the data from the external database. Solving this problem involved data validation, string

matching, and exploring strategies like fuzzy matching and heuristic approaches, concepts I had heard of but had to research more deeply to implement effectively. Exposure to such real-world ambiguity made me appreciate the importance of robust AI pipelines and error-tolerant algorithms.

Looking back, I think that taking more advanced courses in deep learning, transformer architectures, and semantic search would have helped me go beyond rule-based methods and experiment with more scalable, automated approaches to extraction. A course on knowledge graphs or ontology-based systems would also have added value, especially for structuring and linking the extracted entities to external datasets.

This experience has given me a new lens through which I view academic learning. I now better understand how theoretical concepts translate into real applications. My work term did not contradict what I learned in class, but it added depth and context. It reinforced the importance of being able to think critically and creatively when standard tools or algorithms fall short. To better prepare students for such challenges, I think academic programs should offer more interdisciplinary, project-based learning opportunities that mirror real data problems.

Key Learnings and Next Step

This internship helped me understand the challenges of applying NLP in real research settings. Not everything works as expected out-of-the-box, and issues like memory constraints, noisy input data, and mismatches between data sources must be managed creatively.

One surprising insight was how often scientific articles deviate from standardized reporting practices, which affects text extraction reliability. This forced me to explore multiple strategies and balance accuracy with generalizability.

I also realized that clear documentation and code organization are critical especially when collaborating with others who may want to reuse or build upon the code later.

Moving forward, I plan to continue building the classification logic and grouping mechanisms based on trial usage. The lessons I've learned around flexibility, critical thinking, and client-oriented problem solving will be especially helpful in my academic work. I'm confident that the technical groundwork I've laid will support further development and contribute to the research team's broader goals.

Overall, this internship helped me mature as both a programmer and a researcher. I gained firsthand experience in navigating real-world data complexities, adjusted my approach based on evolving requirements, and contributed meaningfully to a research project with real clinical relevance.

Conclusion

This internship has been a transformative experience, bridging the gap between academic learning and real-world application in the field of artificial intelligence and clinical research. Through this project, I deepened my understanding of natural language processing, specifically in the context of processing unstructured biomedical text. I learned how to navigate practical challenges such as inconsistent data formats, ambiguous clinical trial mentions, and the limitations of rule-based systems. This required me to design creative, adaptable solutions and improve the accuracy and scalability of the data pipeline.

The experience also enhanced my collaboration skills. I worked closely with researchers, domain experts, and fellow developers, and learned the importance of clear communication, well documented code, and iterative progress. It taught me how to translate vague goals into concrete deliverables, while staying responsive to evolving project needs.

From an academic perspective, the internship helped reinforce key technical concepts such as text extraction, regular expressions, API integration, and Python based data workflows. I also explored emerging AI techniques including document similarity and contrastive learning, expanding my toolkit beyond traditional methods.

Going forward, I plan to build upon the foundational work completed during this term and explore more advanced NLP and machine learning methods for classification and semantic search. This experience has strengthened both my technical and problem-solving abilities and has reaffirmed my interest in research that combines AI with healthcare. I feel more prepared for future academic challenges and excited about the possibilities that lie ahead.

References

- Hakala, J., & Kekalainen, J. (2017). Automatic classification of clinical trials in medical literature. *Studies in Health Technology and Informatics*, 245, 203–207.
- Brown, T., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. https://arxiv.org/abs/2005.14165
- Gao, T., Fisch, A., & Chen, D. (2021). Making Pre-trained Language Models Better Few-shot Learners. *ACL 2021*. https://arxiv.org/abs/2104.08773
- Schick, T., & Schütze, H. (2021). Exploiting Cloze Questions for Few-Shot Text Classification and Natural Language Understanding. *EMNLP 2021*. https://arxiv.org/abs/2001.07676
- Cheng, J., & Wang, Y. (2019). PDF mining: A survey of document extraction techniques. *ACM Computing Surveys*, 52(4), 1–36. https://doi.org/10.1145/3329782