Text Mining in Rheumatology Research Publication

Archana Senthil; UCID 30265251

The scope of the project is to support the Rheumatology department in automating the identification of clinical trials referenced in research papers, with the broader aim of assisting in the classification and evaluation of relevant scientific literature. This initiative is driven by the need to streamline the literature review process and enhance the reliability and efficiency of linking academic studies to officially registered clinical trials.

In academic and clinical research, determining whether a publication references a known clinical trial is often a labor-intensive task, requiring manual inspection and verification. This project seeks to reduce that dependency by introducing automation in the extraction and matching of trial registration identifiers. The long-term vision includes creating a robust, scalable pipeline capable of integrating with departmental workflows and supporting more advanced applications such as clinical evidence mapping, bibliometric analysis, and systematic reviews.

So far, significant progress has been made in processing research papers for the identification of clinical trials. The contents of the articles have been successfully extracted from a range of data formats including JSONL, CSV, and PDF, ensuring flexibility and compatibility with diverse publication sources. A particular focus was placed on the processing of PDF documents, as these often pose challenges due to their nonstandard formatting. Recognizing that many scientific articles are presented in two column layouts, a custom function was developed to interpret and reconstruct the text in a natural, linear reading order. This functionality ensures that text is semantically preserved and logically sequenced, which is essential for reliable downstream analysis.

In parallel, a comprehensive suite of regular expressions was implemented to detect clinical trial identifiers embedded within the text. These expressions account for the wide variability in how trial IDs are reported, including prefixes, delimiters, and regional codes, ensuring broad coverage of formats such as ISRCTN, EUCTR, CTRI, NTR, RA numbers, and others. The system can capture these identifiers even when they are interspersed with punctuation or embedded in narrative sentences. Moreover, mechanisms have been added to normalize and clean trial IDs prior to comparison, thereby increasing matching precision.

Additionally, the extracted trial identifiers are cross referenced with known clinical registration numbers from external data sources. Merging these references with metadata from structured CSV files allows for automated validation, linking, and classification of papers based on their use of clinical trials. The foundation laid so far enables the integration

of clinical trial identification into a broader framework for scientific literature review, facilitating more efficient and accurate assessment of study provenance and regulatory alignment. This progress represents a key step toward an intelligent and automated pipeline tailored to the needs of the Rheumatology department.

As the next phase of development, the project will focus on organizing the research papers by their publication dates. This will help determine when each clinical trial was first referenced in the literature, providing a clear view of how the use of specific trials has appeared over time. Sorting the papers in this way will also support the analysis of trends in clinical research and show how certain studies have influenced later work.

In addition, we plan to integrate large contrastive language models to help classify the research papers. These models can understand the deeper meaning of the text and can be used to group papers by their focus—such as reporting trial outcomes, reviewing prior research, or exploring new treatment methods. This classification will improve the ability to sort and prioritize the literature for more efficient review.

By combining publication date ordering with automated classification, the system will become more effective in supporting the Rheumatology department's goals. It will not only extract clinical trial identifiers from multiple formats like PDFs, CSVs, and JSONL files, but also help interpret how and why each paper connects to registered trials. This will lead to a more structured, searchable, and meaningful research database.

There are several potential barriers to the current approach that may affect the efficiency and accuracy of automating the identification of clinical trials in research papers. One of the primary challenges is the inconsistent way in which clinical trial identifiers are reported in academic literature. These identifiers can vary in format, may contain typos, or be referenced only in figures, footnotes, or supplementary sections, making it difficult for pattern-based extraction to consistently detect them. Additionally, extracting content from PDFs presents its own complexities, especially when dealing with multi-column layouts or embedded tables, which can result in disorganized or incomplete text retrieval. Access restrictions to some academic papers, such as paywalls or licensing limitations, can also impede the ability to retrieve full texts for analysis. Furthermore, integrating information across diverse file formats like JSONL, CSV, and PDF requires careful handling of schema differences and data type inconsistencies, which can introduce errors if not managed correctly. The reliance on regular expressions for trial ID extraction, while effective in many cases, may struggle to scale as new or previously unseen formats emerge, necessitating ongoing manual updates. A significant limitation also lies in the lack of a large, labeled dataset needed for training and validating classification models, making it difficult to evaluate the accuracy of linking papers to trials or categorizing their content. Finally,

considerations around data privacy and ethical use must be kept in mind, particularly if any of the papers contain sensitive or identifiable clinical data. Addressing these barriers will require a balanced approach that combines technical automation with expert input and iterative refinement.

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