Relation Extraction of Biomedical Text

Project Workbook

By

Neeharika Yeluri (neeharika.yeluri@sjsu.edu)
Pranav Chellagurki(pranav.chellagurki@sjsu.edu)
Rahul Raghava Peela(rahulraghava.peela@sjsu.edu)
Sai Prasanna Kumar Kumaru(saiprasannakumar.kumaru@sjsu.edu)

10/5/2023

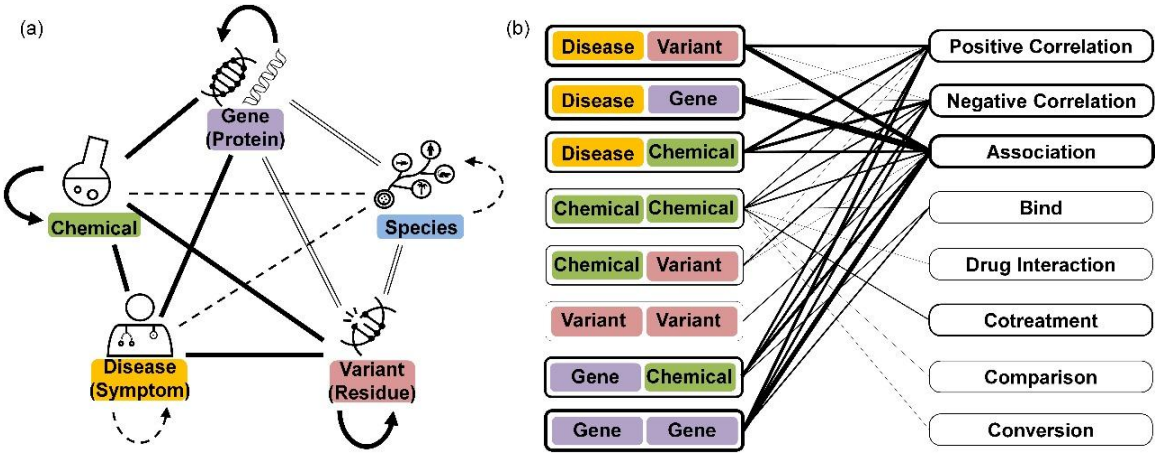
Advisor: Carlos Rojas

Chapter 1. Literature Review

The field of relation extraction has existed for quite some time, especially within the realm of biomedical research. Extracting meaningful relationships from vast and intricate biomedical data is crucial for advancing our understanding of complex biological processes and systems. The drive to discover and discern patterns from what often appears to be chaotic data transcends the preference for any specific model or computational architecture.

In recent times, there has been a noticeable shift towards data-driven approaches in relation to extraction. These techniques, anchored in the paradigm of machine learning, focus on letting the data speak for itself, often leading to more adaptable and robust systems. For our review, we have narrowed down our exploration to such data-driven methods that inherently involve some form of learning.

One such data-driven approach is the development of a novel dataset named BioRED[1], which has been specifically intended to enhance the performance of automated Relation Extraction(RE) and Named Entity Recognition(NER) in the field of biomedicine. The dataset comprises more than 600 papers sourced from PubMed initially and other datasets such as NCBI disease, tmVAR, etc., to be consistent with the previous annotation work. Each abstract has been meticulously annotated among “six commonly described entities (i.e., gene, disease, chemical, variant, species, and cell line) in eight different types (e.g., positive correlation)”[1]. Previously existing corpora were mainly based on the relation between entities in a single sentence. However, the dataset mentioned has annotations that are to the document level i.e., relation among multiple entities among different sentences.



In 2019, Yannis Papanikolaou et al. introduced a groundbreaking approach to relation extraction without direct supervision [6]. Rather than leaning on conventional methods that heavily rely on gold data or distant supervision, Papanikolaou et al.'s technique capitalizes on syntactic parsing and pre-trained word embeddings to discern precise relations. These relationships subsequently inform the annotation of a broader corpus, which is utilized to fine-tune an established BERT model[6]. Through their empirical evaluations of four distinct biomedical datasets, the team showcased the prowess of their methodology, overshadowing two unsupervised baselines and clinching top-tier results in three of the datasets. Notably, the work by Papanikolaou et al[6]. stands out for its ability to adeptly fine-tune a vast pre-trained language model with data that could be deemed noisy.[6]

However, while the method proposed by Papanikolaou and his colleagues presents innovative strides in the realm of relation extraction, it is not devoid of limitations. The technique is bound to sentence-level relations, potentially neglecting more complex interactions that traverse several sentences. The intrinsic reliance on syntactic parsing might introduce inaccuracies and confine its versatility across varied domains or linguistic environments[6]. Furthermore, the approach demands a set of precise relations for the annotation of the larger corpus, a requirement that might pose challenges in certain contexts or languages.[6]

Several prior studies ventured into distant supervision-based biomedical Relation Extraction but contended with the inherent noise in the labeled data. One pivotal challenge was transitioning from sentence-level to corpus-level relation extraction while adeptly mitigating this noise. Addressing this lacuna, the groundbreaking work by Amin et al. (2020) pioneered a novel strategy: extending the conventional sentence-level relation-enriched BERT model to operate at the bag level using multiple instance learning (MIL) techniques [5]. MIL, a paradigm focusing on bags of sentences containing entity pairs, provides a more holistic context, capturing complex relations spanning multiple sentences. Amin and colleagues introduced a sophisticated encoding scheme, notably the k-tag encoding. This directional encoding, employing tags like s-tags for sentence-ordered entities and k-tags for knowledge base (KB)-ordered entities, adds a layer of granularity crucial for distinguishing between different relation types (Amin et al., 2020) [5].

This seminal work significantly advances the landscape of biomedical RE research, emphasizing the pivotal role of data quality in distant supervision. By seamlessly integrating MIL techniques with the power of BERT, Amin et al. demonstrate a substantial leap in accuracy and noise reduction [5]. The experimental validation of two

prominent biomedical relation extraction datasets solidifies the efficacy of their approach (Amin et al., 2020) [5]. This study not only underscores the importance of sophisticated techniques in handling noise but also contributes a practical and efficient methodology for mitigating it in biomedical relation extraction.

Comparison paper

In a comparative study [2] conducted, authors explained how knowledge graphs are essential for representing semantic information in biomedical text extraction. These knowledge graphs provide biomedical literature insights as structured networks of entities and relationships. Rule-based, machine learning-based (Naive Bayes and Random Forests), and transformer-based models (DistilBERT, PubMedBERT, T5, and SciFive) are compared and evaluated in this paper [2]. The rule-based approach, though capable of extracting numerous relationships, exhibits constraints due to its reliance on predefined rules and the constant need for updates to accommodate evolving language nuances. Machine learning methods enhance precision and recall, especially when data balancing is applied, yet exhibit variations in performance across diverse relationship types. Notably, transformer-based models, notably PubMedBERT, shine in this study, achieving an impressive F1-score of 0.92, underscoring their efficacy in biomedical relationship extraction. However, challenges persist, such as dependency on trigger phrases and grammatical completeness, signaling the imperative need for ongoing enhancements.

While the study primarily focuses on the comparison of these relationship extraction methods, an equally vital aspect involves the subsequent step: normalizing entities and relationships. Normalization serves as the foundation for constructing knowledge graphs and aligning entities and their relationships with a standardized vocabulary. Diverse strategies contribute to this normalization process, encompassing the integration of dictionaries, ontologies, and knowledge bases. However, it's essential to acknowledge the intricate nature of this task, characterized by challenges like data quality, entity linking, and relationship normalization. The adoption of transformer-based models, as delineated in the paper, presents a promising avenue to streamline and enhance the efficiency of this normalization process, thereby fortifying the construction of comprehensive biomedical knowledge graphs.

STATE OF THE ART SUMMARY

Scientific Natural Language Processing (NLP) has long grappled with the challenge of acquiring labeled data, a laborious and costly process requiring domain expertise. Addressing this challenge, recent advancements in NLP models have pioneered tailored solutions to elevate the accuracy and efficiency of scientific text analysis. In

2019, Devlin et al. introduced BERT, a groundbreaking language model, which inspired subsequent research in domain-specific adaptations [4]. Building upon BERT's foundation, Beltagy, Lo, and Cohan (2019) pioneered SCIBERT, a specialized language model meticulously honed for scientific texts, explicitly designed to alleviate the scarcity of labeled scientific data in NLP tasks [4].

SCIBERT's genesis lies in its methodical approach to utilizing unsupervised pre-training on a vast corpus of scientific publications, comprising 1.14 million full-text papers from diverse scientific domains. The creation of SCIVOCAB, a specialized vocabulary sourced from scientific texts, marked a pivotal departure from the general domain BERT. This meticulous approach was documented by Beltagy, Lo, and Cohan (2019), and their work set a cornerstone for tailored language models in scientific text analysis.

Upon evaluation, SCIBERT's prowess became evident. The model exhibited remarkable superiority, surpassing BERT-Base by a significant margin. Beltagy, Lo, and Cohan (2019) reported a +2.11 F1 improvement with fine-tuning and an even more impressive +2.43 F1 improvement without fine-tuning [4]. Moreover, SCIBERT emerged as a leader in scientific NLP tasks, achieving new state-of-the-art results in tasks such as Named Entity Recognition (NER), PICO extraction, Text Classification, Relation Classification, and Dependency Parsing. Particularly noteworthy was SCIBERT's exceptional performance in biomedical tasks such as BC5CDR and ChemProt, as well as in EBM-NLP, as highlighted by Beltagy, Lo, and Cohan (2019) [4].

One pioneering model that has garnered considerable attention is BioBERT, introduced by Lee et al. (2019) [7]. This specialized language representation model was meticulously pre-trained on vast datasets from PubMed and PMC, The authors recognized the limitations of employing general-purpose language models in the biomedical domain, where the intricate nuances of biomedical terminology and relationships often elude such models [7]. BioBERT, in this regard, emerged as a compelling solution to bridge this gap, offering the biomedical research community a dedicated tool for enhancing the understanding of biomedical text. Although having an architecture similar to the BERT, BioBERT could outperform previous state-of-the-art models with significant margins in various types of biomedical data mining. As it was trained on the biomedical data. "BioBERT significantly outperforms them on the following three representative biomedical text mining tasks: biomedical named entity recognition (0.62% F1 score improvement), biomedical relation extraction (2.80% F1 score improvement), and biomedical question answering (12.24% MRR improvement)" [7].

A key technical strength of BioBERT is its profound capacity to capture the intricate web of biomedical terminology. Lee et al. (2019) adeptly fine-tuned the model on various biomedical natural language processing (NLP) tasks, including named entity recognition, relation extraction, and question answering, showcasing its superior performance in comparison to its predecessors [7]. We are planning to keep BioBERT as a baseline model to compare and evaluate our model performances.

REFERENCES

[1] Luo, L., Lai, P., Wei, C., Arighi, C. N., & Lu, Z. (2022). BioRED: A Rich Biomedical Relation Extraction Dataset. *ArXiv*. https://doi.org/10.1093/bib/bbac282

[2] Milosevic, N., & Thielemann, W. (2022). Comparison of biomedical relationship extraction methods and models for knowledge graph creation. *ArXiv*. https://doi.org/10.1016/j.websem.2022.100756

[3] Lai, P., Wei, C., Luo, L., Chen, Q., & Lu, Z. (2023). BioREx: Improving Biomedical Relation Extraction by Leveraging Heterogeneous Datasets. *ArXiv*. /abs/2306.11189

[4] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. *ArXiv*. /abs/1903.10676

[5] Amin, S., Dunfield, K. A., Vechkaeva, A., & Neumann, G. (2020). A Data-driven Approach for Noise Reduction in Distantly Supervised Biomedical Relation Extraction. *ArXiv*. /abs/2005.12565

[6] Papanikolaou, Y., Roberts, I., & Pierleoni, A. (2019). Deep Bidirectional Transformers for Relation Extraction without Supervision. *ArXiv*. /abs/1911.00313

[7] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *ArXiv*. https://doi.org/10.1093/bioinformatics/btz682

Chapter 2. Project Justification

In the ever-evolving landscape of Natural Language Processing (NLP), language models, particularly transformer-based ones, have opened new doors of opportunity. Initially trained in general text sources like BookCorpus and English Wikipedia, possess remarkable capabilities for understanding and generating sequential data. However, their application in the intricate realm of biomedical literature poses distinct challenges. Our project is a resolute response to one central challenge: improving the precision and recall of extracting complex relationships between biomedical entities from unstructured textual data.

Biomedical text analysis is the linchpin for extracting invaluable insights from the vast sea of unstructured textual data. Within the realm of biomedical literature lies a treasure trove of intricate relationships between entities such as genes, diseases, drugs, and proteins. Accurate identification and classification of these relationships serve as a catalyst for advancing medical research and expediting drug discovery. While existing methods, including the most advanced approaches like SPINN, BioBERT, and SciBERT, have made commendable strides, the ceaseless evolution of language models and NLP techniques presents a tantalizing opportunity to elevate the precision and relevance of relation extraction within this domain.

Since the advent of BioBERT in 2019, language models have undergone a remarkable metamorphosis. With our project, we endeavor to harness cutting-edge innovations in foundational models, including Large Language Models (LLMs), autoencoders, and traditional recurrent-based networks. Our selection of meticulously curated biomedical text corpora, such as PubMed Abstracts and PMC Full-text articles, underscores our dedication to using high-quality data sources. Through a systematic series of experiments and intensive research, we aim to immerse ourselves in the current landscape of language models, all with the singular purpose of pushing the boundaries of relationship extraction within the biomedical domain.

Our project is not merely an academic exercise but a quest for tangible progress:

1. **Precision and Recall Enhancement**: By capitalizing on the advanced capabilities of language models, our foremost objective is to elevate the precision and recall in identifying and categorizing complex relationships between biomedical entities. This translates to more dependable and valuable insights for the dedicated researchers in this field.

2. **Adaptability and Generalization**: We aspire to demonstrate the adaptability of these models across the multifaceted spectrum of biomedical entities and relationships. The

ability to generalize findings across diverse biomedical subdomains is essential for the widespread applicability of our research.

3. **Scalability and Efficiency**: Our project is also committed to scrutinizing the efficiency of our proposed models when handling extensive volumes of biomedical text data. Scalability is the cornerstone of practical applicability within real-world scenarios.

Our project is an unwavering response to the perennial challenge of precise relationship extraction within biomedical text analysis. The rapid evolution of transformer-based models and NLP techniques offers a tantalizing opportunity to enhance the precision and recall of relation extraction within this specialized domain. Through the infusion of cutting-edge language modeling, we endeavor to contribute significantly to the refinement of tools available to biomedical researchers. Ultimately, our aim is to propel the boundaries of knowledge in biomedicine, empowering groundbreaking discoveries in medical science.

Chapter 3. Project Requirements

Requirements

1. **Essential Features:**
    a. **Datasets Selection:** Identify and acquire relevant biomedical text datasets, such as PubMed Abstracts and PMC Full-text articles, GAD, EU-ADR, and Chemprot.
    b. **Data Preprocessing:** Implement data cleaning, tokenization, and lemmatization processes to prepare the raw textual data for NLP tasks.
    c. **Model Development:** Utilize transformer-based models (BioBERT, SciBERT, PubMed BERT), including Large Language Models (LLMs), autoencoders, and recurrent networks, to develop and train specialized models for biomedical relation extraction.
    d. **Training Pipeline:** Create an efficient training pipeline that includes data loading, model training, validation, and testing stages. Implement mechanisms for model checkpointing and resuming to ensure reliability.
    e. **Evaluation Metrics:** Implement standard evaluation metrics such as precision, recall, F1 score, and accuracy to assess the performance of the developed models accurately.
    f. **NLP Techniques:** Apply advanced NLP techniques such as entity recognition, relation extraction, and semantic role labeling to extract meaningful information from biomedical texts.
2. **Desired Features:**
    a. **Fine-tuning Mechanism:** Implement a fine-tuning mechanism that allows researchers to fine-tune the pre-trained models on domain-specific biomedical data for enhanced performance.
    b. **Transfer Learning:** Explore and implement transfer learning techniques, allowing the models to leverage knowledge from general domain data (e.g., BookCorpus, English Wikipedia) and adapt it for biomedical domain tasks.

Chapter 4. Dependencies and Deliverables

Dependencies

1. **Language Models:**
    a. **BERT-Based Models:** Large Language Models (LLMs), including BERT (Bidirectional Encoder Representations from Transformers) and its variants (BioBERT, SciBERT and PubMed BERT).
    b. **Autoencoders:** Neural network architectures for unsupervised learning, crucial for feature extraction and dimensionality reduction.
    c. **Traditional Recurrent-Based Networks:** Classical models like LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) for sequential data analysis.
2. **Specific Tasks and Tools:**
    a. **Data Preprocessing:** Utilization of Python for data cleaning, text preprocessing, and formatting tasks.
    b. **NLP Libraries:** Integration of NLP libraries such as NLTK (Natural Language Toolkit) for text tokenization, lemmatization, and other language processing tasks.
    c. **Relation Extraction:** Techniques to identify and categorize semantic relationships between biomedical entities, including associations between genes and diseases, drug-target interactions, and protein-protein interactions.
    d. **Evaluation Metrics:** Usage of appropriate metrics (precision, recall, F1 score) to evaluate the performance of language models in biomedical relation extraction tasks.
3. **Datasets:**
    a. **PubMed Abstracts:** Subset of PubMed containing concise biomedical literature abstracts.
    b. **PMC Full-text Articles:** Comprehensive biomedical articles providing in-depth information about various entities and their relationships.
    c. **GAD** (Genetic Association Database)**:** A database of genetic associations with human diseases.
    d. **EU-ADR** (Exploring and Understanding Adverse Drug Reactions)**:** A dataset of electronic health records for drug safety research.
    e. **Chemprot:** A dataset of chemical-protein interactions. These datasets are commonly used in biomedical text-mining research.

Deliverables

Experimental Results Documentation: Detailed documentation of experiments conducted, including methodologies, parameters used, and results obtained. This could include tables, charts, and graphs illustrating the performance metrics of the developed models in comparison to existing benchmarks.

Comparison Report: A comprehensive report comparing the performance of the developed models with existing transformer-based models like BioBERT. Highlight the strengths and weaknesses of each model, showcasing where the new models excel and where improvements are still needed.

Fine-tuning Guidelines: Provide guidelines and best practices for fine-tuning transformer-based models specifically for biomedical text mining. This could include recommendations on dataset selection, preprocessing steps, hyperparameter tuning, and model evaluation strategies.

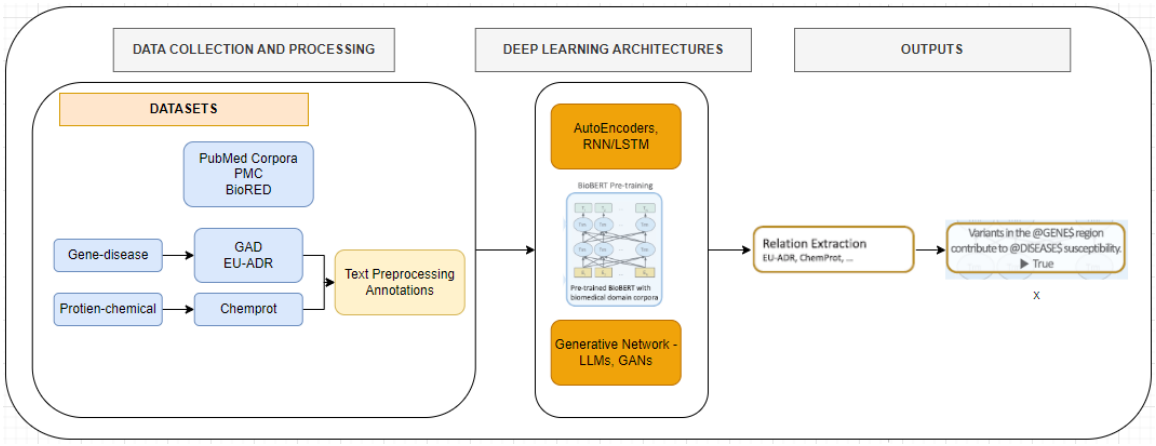
Future Scope: Understanding the types of errors (false positives, false negatives, etc.) and their patterns in reaction extraction can provide valuable insights for future improvements and research directions.

Data Visualization: Visual representations of extracted relationships can aid researchers in comprehending complex networks, thereby enhancing the utility of the developed models.

Technical Documentation: Detailed technical documentation outlining the architecture, algorithms, and technical specifications of the developed models. This documentation would be valuable for researchers and developers interested in understanding the underlying technology.

Derivative Paper for Publication: Prepare a research paper summarizing the methodologies, findings, and contributions of the project. Submit this paper to relevant conferences or journals in the field of biomedical text mining and natural language processing for potential publication.

Chapter 5. Project Architecture



Data Collection:

- Biomedical Text Corpus: Collecting a significant corpus of biomedical text data. Data from sources like PubMed, PubMed Central, etc.
    - Gene-Disease relation Datasets- GAD, EU-ADR
    - Protien-chemical relation Datasets- Chemprot
- Pre-trained Language Models: Consider incorporating pre-trained language models like BioBERT, PubMedBERT, or domain-specific versions of popular models (e.g., SciBERT). These models can serve as a starting point or feature extractor for our architecture.

Data Processing:

- Text Preprocessing: Clean and preprocess the textual data. Mask or replace entity names (e.g., "GENE1," "DRUG2") to help model in recognizing relationships.
- Entity Recognition: Implement entity recognition techniques to identify and normalize biomedical entities within the text.
- Relationship Annotation: Annotate or extract relationships between entities in the text, considering both entity pairs and their specific relationship types.
- Data Split: Divide data into training, validation, and test sets for model development and evaluation.

Model Training:

- Training baseline model on the above processesed dataset.
- Develop and implement the latest NLP architectures such as LLMs, or fine-tune State of the art model architectures with minor modifications in architectures to improve existing models' performance.

Evaluation:

- Model Evaluation: Assess the model's performance on the validation and test datasets. Use the defined evaluation metrics to measure how well the model extracts biomedical relationships.
- Baseline Comparison: Compare the performance of the model against baseline models, including BioBERT or other existing methods, to highlight improvements achieved by our approach.
- Cross-validation: If possible, perform cross-validation to ensure robustness and reduce the risk of overfitting.
- Error Analysis: Conduct error analysis to identify common sources of model errors and areas for potential refinement.

Conclusion and Future Work:

- Interpret Results: Interpret the model's performance and the implications of findings in the context of biomedical relationship extraction.
- Future Directions: Discuss potential future directions for improving the approach, such as leveraging more advanced transformer models or exploring additional data sources.
- Applications: Highlight the practical applications of research, such as aiding in drug discovery, knowledge graph construction, or biomedical information retrieval.

Chapter 6. Evaluation Methodology

Chapter 7. System Design/Methodology

Chapter 8. Implementation Plan and Progress

Project Initiation:

- Thinking of the scope of the project and the use it could serve over the due course of time
- Stating the problem statement in a structured manner i.e., defining a proper end goal to be achieved.
- Having a schedule and team meeting with the professor to guide through the steps to be taken for the project
- Making a project schedule and plan accordingly.

Project Planning:

- Gathering the resources required for the core understanding of the project
- Sorting the appropriate datasets that are related to the project and would be suitable for use.
- Identify the algorithms/models required for the research idea to progress with.
- Also, curating the state-of-the-art for the related research
- Plan for the progress across the semester and towards achieving the end goal such as meeting, milestones, and doubt sessions.

Project Development:

- Gathered research papers related to Relation Abstraction in the Bio-medical field.
- Understand the different terminologies and the progress in the relation extraction to date in the industry.
- Know the tools used to annotate the entity relationship and the datasets created
- Understand the architectures used in the models
- Leverage the latest transformer architectures to improve relation extraction of biomedical text
- Fine-tune and compare results with the state-of-the-art models.

Project Review

- Review the project end result with the goals of the project and compare it with the requirements of the existing requirement in the industry.
- Check for the positive and negative parts of the project and scope for more research.
- Learning to be remembered and drawbacks to be overcome in the future.

Progress

Project Phase	Progress
Initiation	Completed
Planning	Completed
Development	On-going
Review	Scheduled

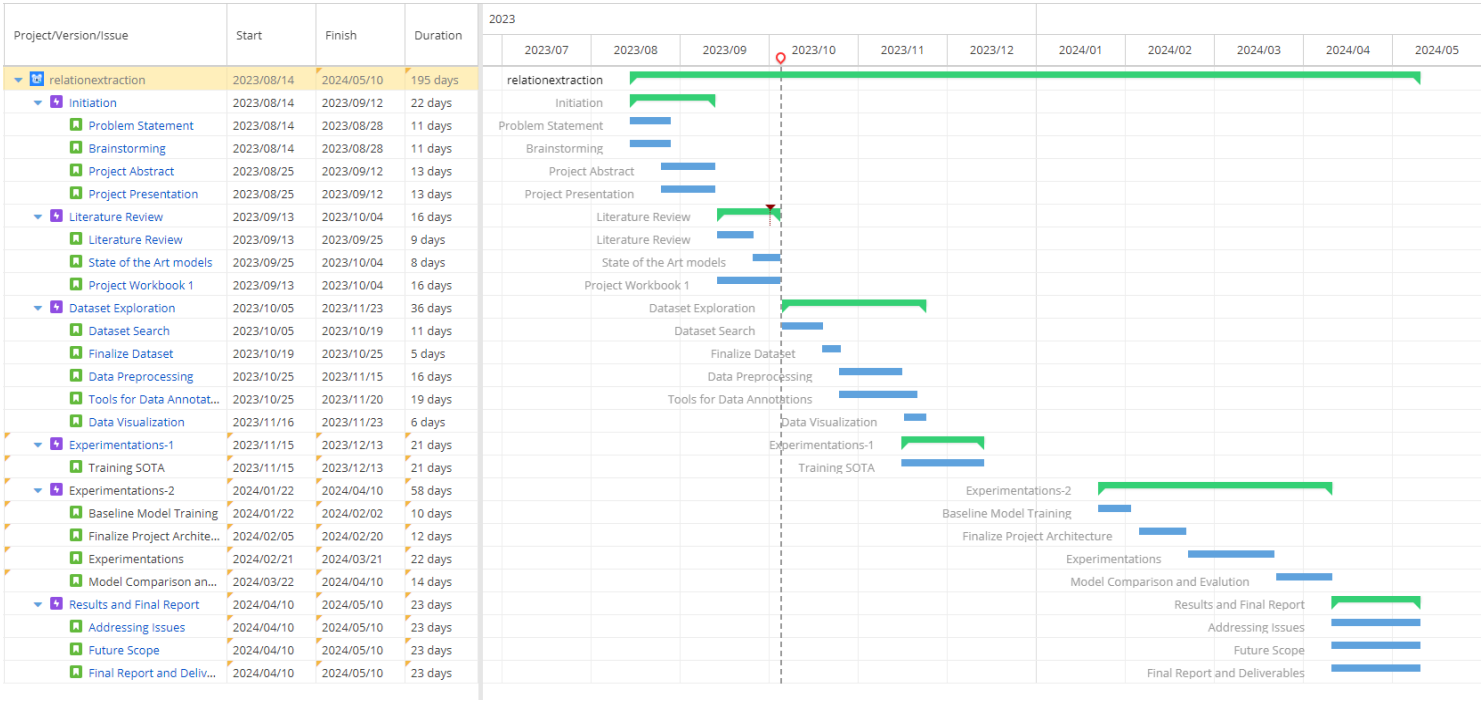
Chapter 9. Project Schedule

Project Schedule

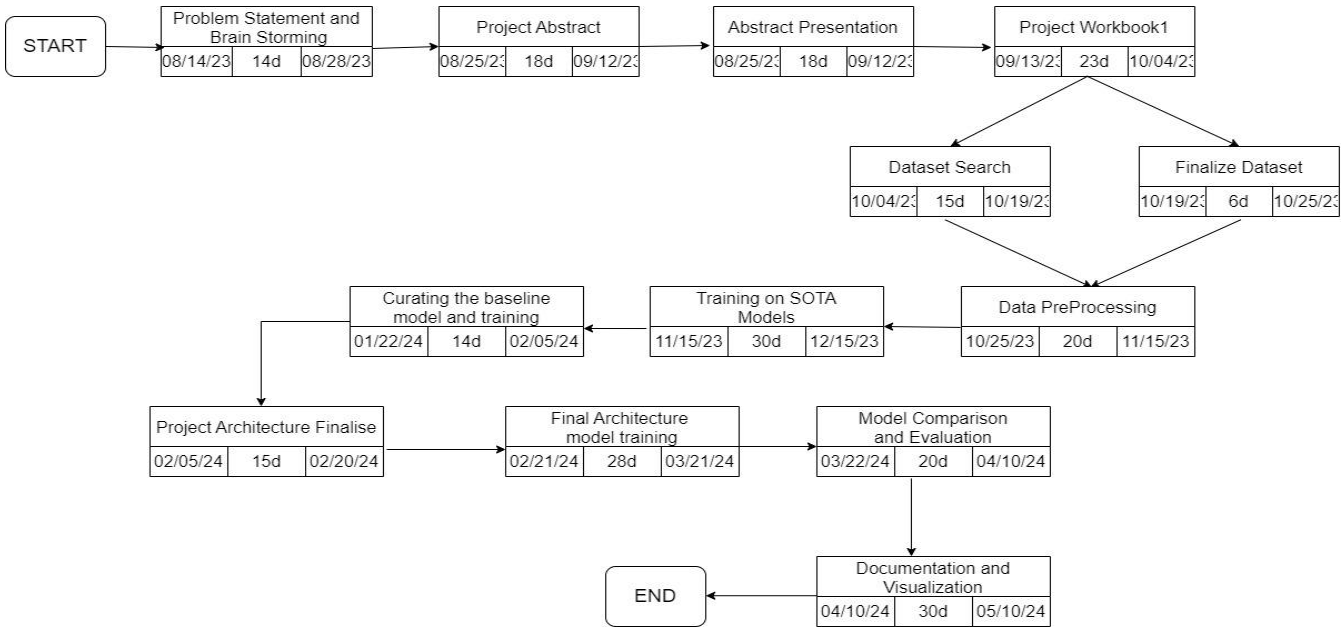
Task	Start Date	End Date	Status	Members
Problem Statement	08/14	08/28	Completed	Team
Brainstorming	08/14	08/28	Completed	Team
Project abstract	08/25	09/12	Completed	Team
Project Presentation 1	08/25	09/12	Completed	Team
Literature review- 1	09/13	10/04	Completed	Team
State of the Art	09/25	10/04	Completed	Team
Project Workbook 1	09/12	10/04	Completed	Team
Dataset Search				
Finalize dataset				
Data Preprocessing				
Tools for Data Annotations				
Data Visualization				
Baseline Model				
Project Architecture Finalize				
Project Workbook 2				
Experimentations				
Model Comparison and Evaluation				
Addressing Issues				
Future Scope				

Final Report and Deliverables				

GANTT CHART



PERT CHART


START
Problem Statement and Brain Storming
08/14/23 14d 08/28/23
Project Abstract
08/25/23 18d 09/12/23
Abstract Presentation
08/25/23 18d 09/12/23
Project Workbook1
09/13/23 23d 10/04/23
Dataset Search
10/04/23 15d 10/19/23
Finalize Dataset
10/19/23 6d 10/25/23
Data PreProcessing
10/25/23 20d 11/15/23
Training on SOTA Models
11/15/23 30d 12/15/23
Curating the baseline model and training
01/22/24 14d 02/05/24
Project Architecture Finalise
02/05/24 15d 02/20/24
Final Architecture model training
02/21/24 28d 03/21/24
Model Comparison and Evaluation
03/22/24 20d 04/10/24
Documentation and Visualization
04/10/24 30d 05/10/24
END