CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Natural Language Inference for Clinical Trials - A Hybrid Approach

by

Junaid Ahmed

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2024

Dedicated to my Parents and Teachers. Whose support and encouragement have been invaluable.

CERTIFICATE OF APPROVAL

Natural Language Inference for Clinical Trials - A Hybrid Approach

by

Junaid Ahmed

(MCS213019)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Malik Ahmad Kamran	COMSATS, Islamabad
(b)	Internal Examiner	Dr. Mohammad Masroor Ahmed	CUST, Islamabad
(c)	Supervisor	Dr. Muhammad Abdul Qadir	CUST, Islamabad

Dr. Muhammad Abdul Qadir
Thesis Supervisor
November, 2024

Dr. Abdul Basit Siddiqui
Head
Dept. of Computer Science
November, 2024

Dr. Muhammad Abdul Qadir
Dean
Faculty of Computing
November, 2024

Author's Declaration

I, **Junaid Ahmed** hereby state that my MS thesis titled "**Natural Language Inference for Clinical Trials - A Hybrid Approach**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Junaid Ahmed)

Registration No: MCS213019

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Natural Language Inference for Clinical Trials - A Hybrid Approach**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Junaid Ahmed)

Registration No: MCS213019

Acknowledgement

I would like to express my deepest gratitude to Almighty Allah, whose unwavering support and guidance have been the foundation of my journey throughout this thesis. Through every challenge and moment of uncertainty, I have felt His presence and strength guiding me. His wisdom and grace have been a constant source of inspiration and perseverance, making this achievement possible. I would like to extend my heartfelt thanks to everyone who has supported and guided me in achieving this milestone. My deepest gratitude goes to Dr. M. Abdul Qadir for shaping the direction of my research, offering invaluable guidance during moments of uncertainty, and encouraging me.

My family has been a major source of strength and encouragement, providing crucial support whenever I needed it. I am deeply thankful to my beloved parents, spouse and siblings for their unwavering encouragement, patience and prayers which have been my greatest source of strength throughout my MS studies.

The academic development I have achieved is largely due to the dedication and support of the professors at Capital University of Science and Technology (CUST), for which I am immensely grateful. Finally, I would like to express my thanks and offer my regards to all those who have supported me in any capacity.

(Junaid Ahmed)

Abstract

Clinical Trials are conducted to evaluate the safety and efficacy of new medications before they are available to the public. The medical domain has experienced rapid growth in research, largely due to the increasing availability of publicly accessible Clinical Trials. These trials are documented in Clinical Trial Reports (CTRs), which include Intervention, Eligibility, Results, and Adverse Events as useful sections. CTRs are typically written in natural language and researchers often rely on them to make inference. A Natural Language Inference (NLI) system for CTRs automatically determines the logical relationship (entailment or contradiction) between a hypothesis and a CTR. This domain has been extensively researched, with around 30 recent articles referenced in this work. The article with the highest score was selected for integration into our hybrid system.

This work presents a hybrid approach to Natural Language Inference on the NLI4CT dataset, utilizing an ensemble system comprising a Multi-Granularity Inference Network (MGNet) and the SciFive model, integrated with rule-based system to enhance the performance of natural language inference system. Rules are formulated after thorough analysis of the training dataset and are subsequently evaluated on the test dataset to assess their effectiveness in improving the accuracy and robustness of the system. Analysis of the dataset revealed recurring patterns and phrases in hypothesis statements, these patterns guided the development of rules for our Rule-based system, which was subsequently applied to the adverse events section.

Our proposed hybrid system outperformed the baseline, achieving an F1 score of 0.870. Rule-based system, despite its simplicity can effectively address specific NLP tasks, especially in domains with well-defined language patterns. Furthermore, Our rule engine identified an annotation error in the NLI4CT dataset. This correction not only improved the quality of the dataset but also contributed to the enhancement of our system's performance.

Contents

List of Figures

List of Tables

Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
cNLP	Clinical NLP
CNN	Convolutional Neural Network
CTRs	Clinical Trials Reports
CoT	Chain-of-Thought
CUST	Capital University of Science and Technology
DL	Deep Learning
EHRs	electronic health records
GLUE	General Language Understanding Evaluation
GAT	Graph Attention Network
GPT	Generative Pre-trained Transformer
LLM	Large Language Model
LPA	Latent Program Algorithm
LoRA	Low-Rank Adaptation
LSTMs	Long Short-Term Memory networks
MGNet	Multi-granularity Inference Network
ML	Machine Learning
MLMs	Masked Language Models
MLRB	ML Pre-processes Input Data for Rule-Based Inference
MTL	Multi-Task Learning
NER	Named-Entity Recognition
NLI	Natural Language Inference
NLI4CT	Natural Language Inference for Clinical Trials

NLP	Natural Language Processing
NLU	Natural Language Understanding
NLM	National Library of Medicine
NIH	National Institutes of Health
PERML	Parallel Ensemble of Rules and ML
QA	Question Answering
RBML	Rule-Based Method Pre-processes Input Data for ML Prediction
regex	Regular Expressions
REML	Rules are Embedded in ML Architecture
RMLT	Rules Influence ML Training
RNNs	Recurrent Neural Networks
RQs	Research Questions
RTE	Recognizing Textual Entailment
SICK	Sentences Involving Compositional Knowledge
SNLI	Stanford Natural Language Inference
T5	Text to Text Transfer Transformer
TREC	Text REtrieval Conference
TE	Textual Entailment
TF-IDF	Term Frequency-Inverse Document Frequency

Chapter 1

Introduction

Information can be expressed in many ways through different sentences that convey similar meanings. This change arises from differences in vocabulary, paraphrasing and writing style . Understanding the semantic similarity between sentences presents a challenge in creating reliable Natural Language Processing (NLP) systems. This challenge has gained significant research interest within the NLP field, especially in the subfield of Natural Language Understanding (NLU). One specific area of NLU designed to address this challenge is NLI, also known as Recognizing Textual Entailment (RTE) [1].

NLP and its subfields have achieved significant advancements in recent years, with their applications in many domains. One such area is the healthcare sector, where there is a need to extract meaningful information from vast amounts of textual data in order to improve clinical research, drug development and patient care. For development of drugs and vaccines, clinical trials are very important. Clinical trials evaluate the safety and effectiveness of new treatments, interventions, and medical devices. The huge amount of information generated during these trials regarding eligibility criteria, adverse events etc. create a challenge for efficient analysis and decision-making. However, huge amount of CTRs are available and more are being published [2].

It is impractical to manually check all the relevant CTRs when finding new treatments. Therefore, an intelligent system is required which can perform the task of reviewing and showing whether claim and available CTRs entails or contradicts.

1.1 Natural Language Inference

NLI is a way to find logical relationship between two sentence, while textual entailment (TE) specifically refers to the task used to predict whether the truth of one text logically follows from another. Recognizing Textual Entailment involves evaluating the semantic relationship between two sentences: the premise, which provides factual information, and the hypothesis, which is assessed to see if its information logically follows from the premise. RTE is crucial in NLP as it helps determine whether two sentences, despite differing vocabulary or syntactic structures, convey similar information [1].

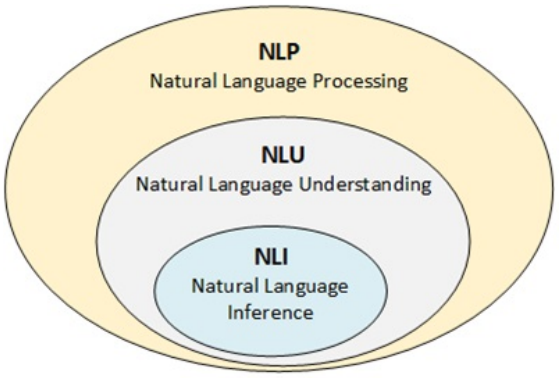


FIGURE 1.1: NLP vs NLI

Goal of NLI is to develop models capable of understanding the logical relationships between natural language statements, specified as premise and hypothesis. The output of NLI system is either Entailment, Contradiction or Neutral. Entailment means that the hypothesis is correct based on premise. i.e. if the premise is true, then the hypothesis must also be true. Contradiction means hypothesis and premise contradicts each other i.e. If the premise is true, then the hypothesis cannot be true and Neutral which represent that the hypothesis is neither entailed by nor contradicts the premise.

Fazelnia et. al. automated requirements engineering tasks using NLI techniques [3]. This work investigates how NLI can be used to extract requirements from natural language descriptions and to find out logical entailments and contradictions between requirements. Landscape of research in NLI has expanded as new NLI datasets and benchmarks are being developed, which can be seen in survey paper by Gubelmann et. al. [4]. This research categorizes existing NLI datasets based on the types of logical inferences they target, and also introduces two novel NLI benchmarks. Research at University of Manchester propose use of scientific explanations to improve explanation-based NLI, which aims to generate natural language explanations for the inferences made by NLI models [5].

TABLE 1.1: An Example of Natural Language Inference

Type	**Statement**	**Relationship**
Premise	A kid is in the park playing baseball	
Hypothesis 1	Child plays baseball with friends	Entailment
Hypothesis 2	A person is waiting to be served his food	Neutral
Hypothesis 3	A kid is playing basketball with friends	Contradiction

Overall, the field of NLI is seeing significant progress in areas such as dataset development, model architectures, and real-world applications, as evident from the research papers cited [3–5]. These developments are helping to advance the state-of-the-art in natural language understanding and reasoning.

Some applications of NLI includes

Question Answering

Which helps in understanding if a given answer is supported by the provided context or passage.

Text Summarization

Assists in generating summaries that are consistent with the original content.

Information Retrieval

Enhances search engines by understanding and matching queries with relevant documents.

1.1.1 CTR Entailments

Entailment refers to a type of semantic relation that holds between sentences, where one sentence logically follows from another. This is an important concept in the context of CTRs. With entailment, it is possible to digitally analyze 1000s of CTR documents without need to manually scan all of these CTRs. Implementing Entailment in CTR can help advancing medical research and practice.

Example

There are three different ways to show textual entailment, examples of which are as under

1. **Positive TE (where the text supports the hypothesis) is:**
    - **Text:** Inclusion Criteria: HER2-positive T1 histologically confirmed invasive carcinoma of the breast
    - **Hypothesis:** Patient suffering from ErbB2+ breast cancer are eligible
2. **Negative TE (where the text contradicts the hypothesis) is:**
    - **Text:** Inclusion Criteria: HER2-positive T1 histologically confirmed invasive carcinoma of the breast
    - **Hypothesis:** Patient suffering from ErbB2+ breast cancer are not eligible
3. **Non TE (where the text neither supports nor contradicts the hypothesis) is:**
    - **Text:** Inclusion Criteria: HER2-positive T1 histologically confirmed invasive carcinoma of the breast
    - **Hypothesis:** Patients suffering from hypertension are more likely to experience organ damage

ClinicalTrials.gov is a detailed and publicly accessible database of clinical trials conducted around the world. It is maintained by the U.S. National Library of Medicine (NLM) at the National Institutes of Health (NIH). The database gives comprehensive information regarding clinical studies to help researchers, health-care professionals, and the public find relevant information on clinical trials. As of 2024-08-30 ClinicalTrials.gov has registered 507,578 studies [6].

1.1.2 Types of NLI Systems

Ensemble System

Ensemble methods rely on the principle of collective decision-making, where a group of individual classifiers collaborate to determine the most appropriate output. This collective decision can be achieved through techniques such as voting or averaging probabilities [7]. Despite the variety of deep learning architectures and their capability to handle complex problems while automatically extracting features, a significant challenge in deep learning lies in the need for extensive expertise and experience to effectively tune the optimal hyperparameters, making this process both tedious and time-consuming [8]. Ensemble learning methods train multiple base learners and combine their predictions to obtain improved performance and better generalization ability than the individual base learners [9]. The integration of hybrid and ensemble techniques in NLP seeks to improve performance metrics across tasks by harnessing the complementary strengths of different models. Ensemble methods can bolster the generalization capabilities of deep learning models, leading to more consistent performance across a range of datasets and domains [10].

Some NLP systems seen in recent literature that uses ensemble methods are: Kim et al. [11] stacked ensemble using a search-based structured prediction algorithm. Alekhya et al. [12] An ensemble methodology for applying natural language processing in healthcare diagnosis. Whereas Baradaran et al. [13] introduced a

method based on ensemble learning aimed at boosting the generalization capabilities of machine reading comprehension systems. [14] introduced an ensemble deep learning strategy optimized for automatic hate speech detection via natural language processing. Zhou et al. [15] introduced a novel weighted ensemble model method for recognizing text implications. Jaradat et al. [16] introduces a novel hard voting classifier designed to improve crash severity classification by integrating machine learning and deep learning models with a range of word embedding techniques, such as BERT, RoBERTa, Word2Vec, and Term Frequency - Inverse Document Frequency (TF-IDF). Fattahi et al. [17] presented SpaML, A bimodal ensemble learning spam detection system utilizing natural language processing techniques. Ansari et al. [18] focuses on training text classifiers for detecting depression, with the primary goal of enhancing the performance of depression detection.

Deep Learning (DL)

Deep learning approaches have been highly successful in addressing the task of Natural Language Inference. Deep learning techniques are surpassing traditional machine learning methods, allowing computational models to progressively learn features from data across multiple levels [19]. Commonly used methods are recurrent neural networks (RNNs) [20], Bidirectional Encoder Representations from Transformers (BERT) [21] , Generative Pre-trained Transformer (GPT) [22] and Long Short-Term Memory networks (LSTMs) [23]. The current state-of-the-art methods for NLI rely on fine-tuning pre-trained neural language models, such as BERT [24].

A significant advantage of these neural methods lies in their ability to generalize across diverse NLI datasets, offering strong performance even when applied to new or challenging examples. This adaptability is particularly useful for complex language understanding tasks, as the models effectively learn to distinguish between different types of relationships through exposure to large datasets and

sophisticated training processes. The key advantages of these neural models are their ability to capture complex semantic and logical relationships from the text, as well as their strong generalization performance across different NLI datasets [25].

However, purely neural network based approaches also have some limitations. They can struggle with tasks that require more explicit logical reasoning, such as handling negation, quantifiers, and compositional semantics. To address this, recent work has explored combining neural methods with symbolic logical reasoning, in frameworks like NeuralLog [26].

Machine Learning (ML)

Machine learning models typically rely on traditional statistical techniques and algorithms such as linear regression, decision trees, or support vector machines. It has excelled in pattern recognition tasks across diverse domains, including computer vision, speech recognition, natural language processing, and game AI [27]. ML algorithms have significantly advanced drug discovery. Pharmaceutical companies have greatly benefited from the utilization of various ML algorithms in drug discovery. ML algorithms have been used to develop various models for predicting different characteristics of compounds in drug discovery [28]. These models require carefully engineered features and can perform well on NLI tasks with relatively small amounts of training data. However, their performance may be limited by the quality and relevance of the engineered features [29].

Rule-based System

Rule-based method is the precursors to modern NLP techniques. In this approach, predefined linguistic rules are used to analyze and process textual data. Rule-based expert systems can mimic human decision-making by leveraging encoded expert knowledge. This knowledge can be acquired from experts or extracted from

various sources. These systems offer several advantages, including permanence and reliability [30].

Rule-based expert systems elevate the quality of traditional decision-making processes by offering visual, transparent, and precise comparisons between observed and expected values. Additionally, they facilitate problem identification and can contribute to process standardization and waste minimization [31].

Rules are often implemented using regular expressions (regex), which proved to be rigid and inflexible, requiring hard-coded rules to match but are well-suited for exact pattern matching [32]. Gabud et al. states rule-based methods are well-suited for relation extraction tasks, where the rules can encode patterns that indicate semantic relationships between entities in the text [33]. The key advantages of rule-based methods are their interpretability and ability to capture domain-specific patterns.

The literature review highlights the application and importance of rule-based systems. In healthcare, rule-based system can provide valuable recommendations or alerts, aiding healthcare professionals in decision-making [34]. QuARS is a Rule-based system developed by the researchers' lab that employs deterministic rules to automatically identify defects in plain text requirements documents. While QuARS effectively performs lexical and syntactical analysis, it currently lacks semantic analysis capabilities [35]. Another example is predicting the risk of virus spread. By employing linguistic terms like "low," "moderate," and "high," these systems effectively capture the inherent uncertainty associated with such predictions [36]. Despite their strengths, rule-based approaches also have limitations.

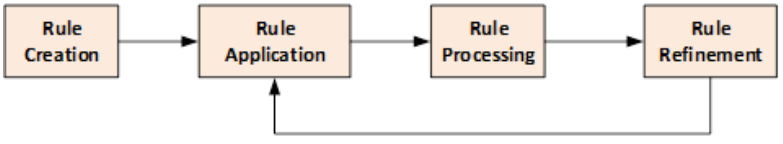


FIGURE 1.2: Steps Involved in Rule-based Method

Creating and maintaining comprehensive rule requires deep linguistic expertise. While rule-based approaches can achieve high precision, they can suffer from lower

recall as they may miss out on patterns not covered by the predefined rules. To address this limitation, hybrid approaches have been proposed that uses rule-based methods along with machine learning or deep learning models to benefit the strengths of both [33, 37].

Cust NER

CustNER is a rule-based NER which is developed by a PhD Scholar of Capital University of Science and Technology (CUST). It is a system for named-entity recognition (NER) that can identify person, organization and location in English text. It uses rule-based approach, which relies on a set of predefined rules and patterns to detect and classify named entities, rather than using deep learning models. Research conducted by Rabia et al. highlights that rule-based systems, especially those with well-defined, customized rules and patterns, can provide significant advantages over more generalized machine learning approaches in specific domains of NLP. When carefully tailored, these rule-based systems excel in tasks like NER, where domain-specific terminology and structured knowledge are essential for accuracy [38].

Hybrid system

A hybrid system refers to a computational framework that combines different methodologies or approaches to leverage their respective strengths and achieve improved performance or capabilities.Despite advancements in Natural Language Inference using large-scale deep models, these models struggle to generalize effectively on adversarial datasets that present complex linguistic challenges. Such datasets often introduce complex ambiguities, and nuanced language phenomena that expose the limitations of purely data-driven approaches. These limitations highlight the need for hybrid systems that combine the strengths of multiple approaches to achieve more robust performance. Which raise need of Hybrid systems [39]. Hybrid systems are based on the integration two or more approaches such as

integration of deep learning with machine learning, rule based, ontology based or some other approach.

Consider rule-based with deep learning approaches for NLI, the combination may leverage the strengths of both methodologies [40]. Rule-based systems excel at accurate and efficient numerical inference on well-defined problems [31], while deep learning models are better equipped to handle complex, ambiguous, and dynamic numerical reasoning tasks [39]. Integrating rule-based numerical reasoning components into a deep learning NLI architecture can improve overall performance of the system [40].

1.1.3 Numerical Inference in Deep learning - A Key Challenge

NLI4CT, requires complex numerical inference in textual entailment task. SOTA deep learning models being best performer for many NLP tasks has shown low performance on numerical inference tasks [41]. Because of this the key challenge in deep learning is using it for numerical reasoning in natural language inference. Current NLI models often struggle with numerical reasoning tasks.

Handling complex numerical reasoning is another challenge.NLI tasks can involve various types of numerical reasoning, such as arithmetic operations, comparisons, and quantitative reasoning as depicted in figure 1.3. Developing deep learning models that can accurately perform these diverse numerical reasoning skills remains a significant challenge. Furthermore, NLI models trained on specific numerical reasoning tasks may struggle to generalize their skills to numerical reasoning problems or subtle variations. Improving the generalization in NLI is an important research direction [42].

Rule-based systems excel at numerical inference tasks that can be limited in number and well-defined through a set of clear rules and parameters. Their deterministic nature and simplicity make them a suitable choice for applications that

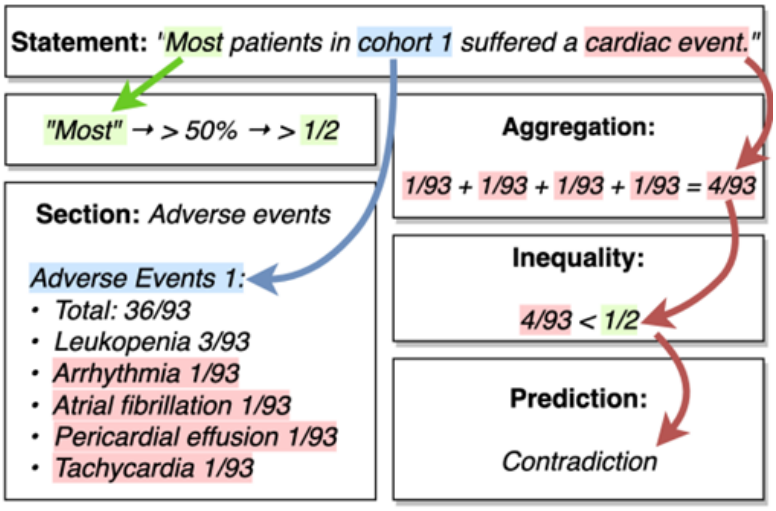
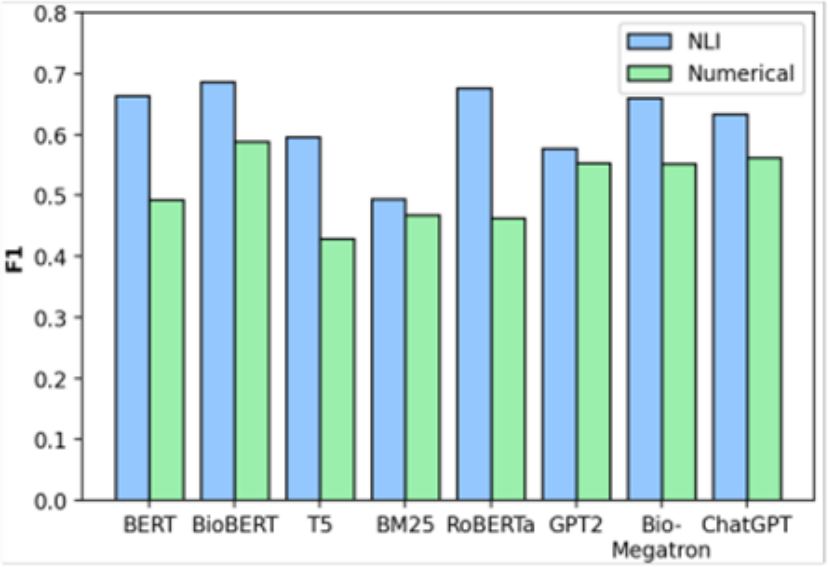


FIGURE 1.3: Schematic of a Numerical Inference Chain to Solve an NLI4CT Instance. Figure from Jullien et. al. [41]



(a) **NLI vs Numerical**

FIGURE 1.4: Comparison of NLI and Numerical Categories on Deep Learning Models. Figure from Jullien et. al. [41]

prioritize speed, accuracy, and ease of implementation over the need for continuous learning and adaptability. Jullien et. al. [41] performed comparison of state of the art Deep Learning models and on NLI and Numerical categories of statements and found that performance of Deep learning based systems on numerical category is less than performance on NLI category as shown in figure 1.4.

1.2 Dataset

Datasets for NLI play a crucial role in advancing the field by providing diverse and challenging sentence pairs that help train and evaluate models on understanding logical relationships between statements. They challenge models to accurately identify entailment, contradiction, and neutrality between statements. The Stanford Natural Language Inference (SNLI) dataset is a large-scale collection of human-written sentence pairs labeled as entailment, contradiction, or neutral, designed to facilitate the development and evaluation of natural language inference models [43].

MNLI is an extension of SNLI, which includes 433,000 sentence pairs across different genres (e.g., fiction, government, telephone conversations). It provides more diverse contexts for NLI tasks [44]. Natural Questions is developed by Google, includes questions paired with passages from Wikipedia. It aims to evaluate models on question answering and entailment tasks [45]. Other datasets for NLI task includes Quora Question Pairs [46], General Language Understanding Evaluation (GLUE) [47], SNLI-VE (Visual Entailment) [48], Sentences Involving Compositional Knowledge (SICK) [49], Text REtrieval Conference (TREC) [50], DocNLI [51] and ContractNLI [52].

The NLI4CT [41] is high quality annotated dataset created for challenging problems of semantic analysis such as Textual entailment. It is aimed to entail hypotheses based on CTRs and retrieve the corresponding evidence supporting the justification. This task is based on a collection of breast cancer CTRs (extracted from https://clinicaltrials.gov/ct2/home), statements, explanations, and labels annotated by domain expert annotators.

This dataset contins 2,400 annotated statements containing CTRs and evidence. A split of 1,700 training samples and test dataset containing 500 entries each is performed on dataset. In order to study the consistency, robustness and faithfulness of large language models in clinical natural language inference we constructed

dataset which contained interventions that focus on numerical reasoning, vocabulary and syntax, and semantics.

FIGURE 1.5: An Example of Clinical Trial Report along with Input Statement and Label. Figure from SemEval, 2023 [53]

One of the goals with SemEval-2023 Task 7 is to generate hypotheses using queries on CTRs and find evidence for why that reason happened. The task: A collection of breast cancer CTRs and statements, paired with explanations and labels annotated by domain expert annotators. Some illustrative examples of CTRs from the NLI4CT dataset are provided in Appendix A, showcasing the structured annotations and inferences necessary for accurate model predictions.

In NLI4CT dataset the Clinical Trials are divided into 4 sections

- **Eligibility Criteria:** - Conditions for someone to be able or allowed into a clinical trial, specific criteria must be met.
- **Intervention:** - Any details relating to the type, dose (inputted in quantity form), timing and duration of treatment being investigated.
- **Results:** - Indicates participants count in the trial, outcome measures, units and results.

- **Adverse Events:** - Diseases and symptoms found in patients during the clinical trial.

Result

Zhou et al. [54] achieved highest performance on NLI4CT dataset. with an overall F1-scores of 0.856 on the which is the highest value achieved. Zhou et. al. used ensembling system by integrating Multi-granularity Inference Network. This system is a pipeline system that models the tasks of evidence gathering and natural language processing separately and SciFive Model which is an advance version of T5 Model developed by Google to resolve the problem of Complex Numerical Inference in NLI4CT Dataset.

Work @Team name	Approach	Generative/ Discriminative	Retrieval type	Pre-training Datasets	F1
(Zhou et al., 2023) @THiFLY	MGNet, BiLSTM and SciFive model ensembling	G + D	Post	PubMed Abstract, PMC	0.856
(Kanakarajan and Sankarasubbu, 2023) @Saama AI Research	Instruction-finetuned LLMs, Flan-T5	G + D	-	-	0.834
(Vladika and Matthes, 2023) @Sebis	Ensemble of a pipeline and joint system based on DeBERTa-v3	D	Pre	-	0.798

FIGURE 1.6: Performance of Leading Systems. Figure from Jullien et. al. [2]

Performance of leading systems on Textual Entailment for NLI4CT dataset is shown in Figure 1.6

1.3 Problem Formulation with Significance

Clinical trial data is usually contains complex information such as intricate numerical data and specialized medical terminology. Analysis of this data is a challenging task. Therefore, improvements in trial analysis methods are needed to accelerate drug development [55]. Leading deep learning system using MGNet and SciFive Models face challenges, particularly with numerical inference, due to the need to accurately interpret and contextualize quantities, percentages, and rates alongside

medical terms. This numeric data is inevitable, the importance of numbers in clinical text is highlighted by Mahendra et al. [56]. Although the leading system on the NLI4CT dataset achieved an F1 score of 0.856 using advanced models for textual entailment, gaps remain in handling nuanced numeric and domain-specific inferences [54].

This highlights the ongoing need for enhanced models that can reliably interpret clinical language and improve the performance and reliability of Natural Language Inference in healthcare contexts. Automated analysis of large volumes of clinical trial data is of great importance, as clinical trial data is inherently complex, containing a mix of structured numerical data, medical terminologies, and unstructured text.

Existing systems are still struggling with effectively handling numerical inference, a key challenge in clinical data interpretation. Clinical Trial data is complex as it contains numerical data and medical terminologies, existing systems are still struggling with numerical inference. The leading system of NLI4CT dataset achieved an F1 score of 0.856 with the best model for textual entailment - [54]. The failure cases of the system include

- Related and similar terms such as 'primary trial', 'cohort' etc. appear frequently across the dataset.
- Simple Numerical expressions like 'less than 5%', '15/200', 'more than half' etc.
- Adverse Event or medical condition names such as 'Fever', 'Death', 'Low platelet count' etc.

Keeping above facts in view, a technique is therefore needed, which is able to tag similar words, numerical expressions & adverse events and perform numerical inference between statements and clinical trial data. Such a technique would bridge the gap between textual understanding and quantitative reasoning, addressing challenges unique to the clinical domain.

1.4 Research Questions

To solve the above mentioned problem, the following research questions (RQs) need to be answered:

- RQ1: What are the shortcomings in the existing top performing NLI system and reasons?
- RQ2: How to formulate a hybrid NLI system based on identified weaknesses in existing system?

1.5 Research Objectives with Scope

- RO1: Find cases in adverse events section where existing SOTA system failed and analyze short comings to design a better system
- RO2: Devise a technique for correct prediction of entailment in Clinical Trial data which are wrongly predicted by the existing SOTA NLI system.

1.6 Research Methodology

Below is the explanation of research methodology presented in Figure 1.7. The methodology provides a structured approach to finding solution for problem.

1. **Identify Problem & Formulate Research Question**
    - The first step in research methodology is to identify the problem.
2. **Understand the Problem**
    - It involves understanding of failure cases from existing SOTA system and shortcomings which led to failure. This is a crucial step, as the entire research is based on it.

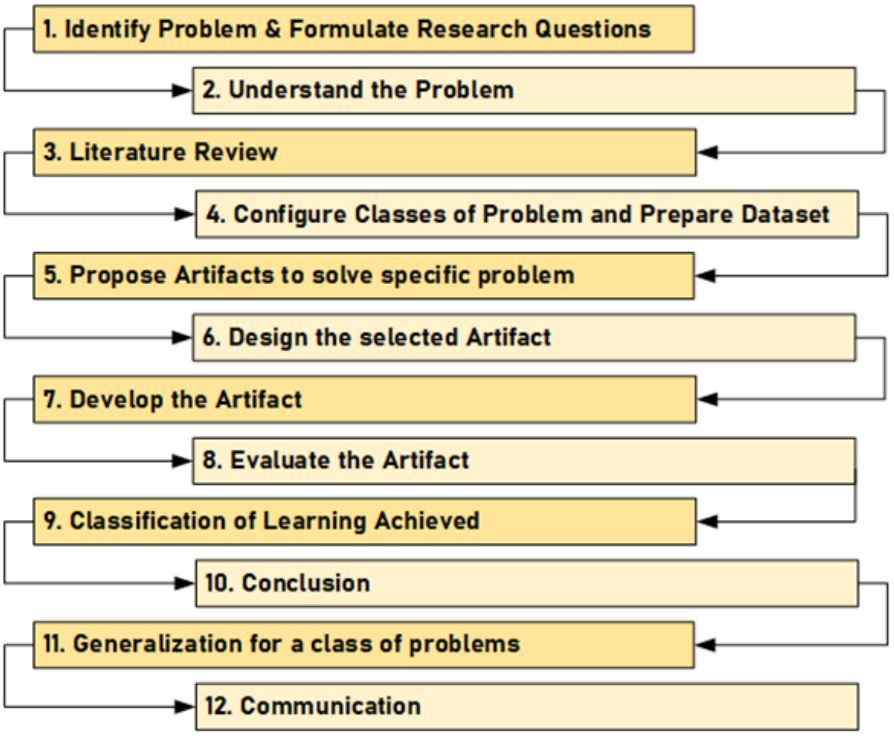


FIGURE 1.7: Steps Involved in Research Methodology

3. **Literature Review**

   - It involves critical analysis of existing publications relevant to a particular topic or research question. This process aims to provide a comprehensive overview and understanding of what is currently known about the subject and to identify gaps, inconsistencies, or areas where further research is needed. In chapter 2, we have discussed a detailed summary of existing literature.

4. **Configure Classes of Problem and Prepare Dataset**

   - In this step, dictionaries are required to tag keywords, numerical expressions, and adverse event names. Furthermore, the NLI4CT dataset needs to be preprocessed in order to apply rules.

5. **Proposed Artifacts to Solve Specific Problem**

   - To address the challenges of accurately determining textual entailment within complex datasets, this study proposes a hybrid Natural Language Inference (NLI) system.

6. **Design the Selected Artifact**

    - Rules will be crafted using training data such that they cover failure cases from the existing SOTA system.

7. **Develop the Artifact**

    - Develop a rule engine to predict whether the statement and clinical trial data entail or contradict.

8. **Evaluate the Artifact**

    - Evaluate the system using Precision, Recall, and F1 Score on evaluation datasets and compare results with the existing system.

9. **Classification of Learning Achieved**

    - Factors identified which have positively contributed to our research success. Identify cases where our system fails.

10. **Conclusion**

    - It involves the culmination of the study, summarizing the key findings, implications, and insights derived from the investigation.

11. **Generalization for a Class of Problem**

    - The rule-based NLI system will find entailment or contradiction from the input pair.

12. **Communication**

    - Methods and results will be communicated via research publication.

Chapter 2

Literature Review

2.1 Introduction

Research in Natural Language Processing within clinical trials is advancing rapidly, leveraging the vast amounts of unstructured data available in clinical trial reports. By integrating NLP with other AI-driven approaches, the goal is to reduce the time and cost associated with clinical trials while ensuring higher accuracy and enhancing patient care through predictive analytics. There are several publicly available Clinical Trials datasets designed.

- Treatment, diagnosis, and side effects extraction from medical text.
- Cohort selection for clinical trials. [57]
- Extracting Participants, Interventions, Comparisons and Outcomes elements from clinical trial abstracts. [58]
- Review of publicly available clinical NLP tasks and datasets. [59]

The TREC 2021 Clinical Trials Track [60] is among one of the initiatives that focused on improving the matching of patients descriptions to eligible CT Reports

TABLE 2.1: Year wise Distribution of Papers Reviewed

Method	**2019**	**2020**	**2021**	**2022**	**2023**	**2024**	**Total**
Ensembler Systems					4		4
Deep Learning based	1		1	1	11	5	19
Rule based		1					1
Hybrid					1		1
Other Methods	1				2	1	4
Total	2	1	1	1	18	6	29

from the ClinicalTrials.gov repository. This track utilized a dataset containing 75 synthetic patient descriptions (5-10 sentences) mimicking admission notes, created by medical professionals. These descriptions were paired with a vast collection of 375,580 clinical trial XML files from ClinicalTrials.gov, which included details like eligibility criteria, conditions, and keywords. Experts then judged the relevance of each patient-trial pair as "eligible" (meeting inclusion criteria), "excludes" (meeting both inclusion and exclusion criteria), or "not relevant". While the TREC dataset lacked extensive labeled training data, participants used external sources like the SIGIR dataset to train their models. Ultimately, the goal was to develop automated systems that could effectively match patients to appropriate trials, leading to improved recruitment and access to new treatments [61].

DeYoung et. al. [62] introduced Evidence Inference 2.0 which is a Question Answering(QA) task and span selection tasks where given an outcome, intervention and comparator intervention, systems need to predict whether the intervention increased the measurement of that outcome significantly, decreased it significantly or no change based on comparison with the comparator. The authors collected additional annotations to expand the Evidence Inference dataset by 25%, resulting in a total of 2,400 statements and comparative treatment relationships. They provide stronger baseline models for two main tasks: inferring the comparative performance of two treatments from a given article, and identifying whether a given statement is supported or contradicted by a CTR.

Romanov [63] presented MEDNLI dataset which involves determining whether a short medical history note supports or contradicts a given statement. In other words, it aims to establish a logical relationship between the two pieces of text.

It also highlights challenges in applying NLI to clinical text due to factors like ambiguity, specialized terminology, and complex reasoning required to understand clinical narratives.

Jullien et. al. [2] presents a novel benchmark called NLI4CT for natural language inference tasks on clinical trial reports. CTRs contain crucial information for developing personalized medicine. All of the above mentioned tasks Previous NLP tasks on CTRs have limitations. They often concentrate on specific CTR sections and involve repetitive reasoning patterns, like checking eligibility or comparing measurements. Unlike these tasks, NLI4CT requires understanding and reasoning across the entire CTR, combining medical and numerical information. The NLI4CT format avoids repetitive reasoning patterns by presenting diverse and complex scenarios. The NLI4CT benchmark includes two main tasks which are determining the inference relation (entailment, contradiction, or neutral) between a natural language statement & CTR and secondly, retrieving supporting facts from the CTR to justify the predicted inference relation.

The authors release a corpus of 2,400 expert-annotated statement-CTR pairs with labels and supporting evidence. The study evaluated six existing NLI models on this dataset, achieving a maximum F1 score of just 0.627. This highlights the limitations of existing NLI approaches, which struggle with the numerical reasoning, domain-specific terminology, and long-form text required for these clinical inference tasks. To the authors' knowledge, this is the first task that covers the full scope of CTRs, combining the challenges of biomedical and numerical natural language inference. The NLI4CT dataset, leaderboard, and code are publicly available to spur further advancements in this important area [41].

Kierner et. al. [64] presented taxonomy of hybrid architectures in clinical decision systems that combine rule-based reasoning and machine learning. Hybrid architectures in clinical decision systems often utilize a combination of rule-based reasoning and machine learning techniques to improve decision-making performance. Five unique architectural types were identified by the Author: Rules are

Embedded in ML architecture (REML), ML pre-processes input data for Rule-Based inference (MLRB), Rule-Based method pre-processes input data for ML prediction (RBML), Rules influence ML training (RMLT), Parallel Ensemble of Rules and ML (PERML). One example of a hybrid approach is a text categorization method that combines a machine learning algorithm with a rule-based expert system. Author indicate that the integration of rule-based reasoning and machine learning leverages the strengths of both approaches and improve overall decision-making performance.

Ray and Chakrabarti [65] proposes a hybrid approach that combines deep learning and rule-based methods to enhance aspect-level sentiment analysis. The researchers identify limitations of existing machine learning techniques in extracting implicit aspects and capturing user sentiment accurately.

To address these issues, the authors use a seven-layer deep convolutional neural network (CNN) to tag aspects in opinionated sentences. They then combine this deep learning approach with a set of rule-based methods to improve both aspect extraction and sentiment scoring performance. The proposed method also incorporates aspect categorization using a predefined set of aspect categories and a clustering technique. Evaluation results show that the overall accuracy of the authors' mixed approach is 87%, which is 7-12% higher than state-of-the-art methods like modified rule-based and standalone CNN models

Chen et. al. [66] developed a rule-based system and a general clinical NLP (cNLP) system. The rule-based system used a flexible framework combined with specialized medical knowledge. The cNLP system relied on the Unified Medical Language System and Unstructured Information Management Architecture. The multi-level, rule-based approach allows the NLP system to handle the complexity and variability of clinical data, which is important for effective cohort selection. Results show that the rule-based system performed exceptionally well in the 2018 n2c2 challenge, ranking fourth out of all participants. While the cNLP system didn't match this level of accuracy, it demonstrated potential for extracting clinical information. Author states, findings show that a carefully designed rule-based NLP

system can effectively identify eligible patients for clinical trials, even with limited data. Additionally, combining rule-based and cNLP approaches could lead to even better results in the future.

Wang et. al. [67] aims to develop a rule-based text classification algorithm to automatically summarize a patient's periodontal health status from unstructured electronic health records (EHRs). The text preprocessing step involves identifying and replacing 170 common medical acronyms with their full forms to improve the accuracy of the algorithm. The proposed approach uses three main steps: text preprocessing, seed word preparation, and keyword extraction. The keyword extraction utilizes an SPPMI-based technique. This work is part of a broader effort to develop robust text mining capabilities based on AI/ML techniques, such as text clustering and keyword extraction, to capture and monitor malicious communication on social media platforms. The research highlights the need for quantitative tools to support human expert intervention in monitoring the increasing volume and velocity of malicious information that may be amplified by the integration of generative AI into social media. The paper presents preliminary steps towards characterizing and monitoring this emerging challenge, including the potential risks posed by the use of generative AI by politically-oriented entities.

2.2 Literature Review

The analysis of existing literature reveals that the majority of the Natural Language Processing systems can be categorized into three primary methods: Ensembling Systems, Deep Learning based systems and other methods. Where majority of work is being done on deep learning techniques esp. transformer models.

2.2.1 Ensemble Systems

In recent studies, various methodologies have been explored to enhance model performance in multi-task learning within the clinical NLI domain. Ensemble

techniques where multiple models are combined to achieve better performance have been seen in recent literature.

System developed by Zhou et. al. [54] received the honor of the top-scoring systems and gained significant performance increase in Task 1 and Task 2. The author proposes a multi-granularity inference system to address the challenges of the A Multi-evidence Natural Language Inference for Clinical Trial Data task of NLI4CT. Therefore, this involves formulating hypotheses of the task based on Clinical Trial Reports and extracting supporting evidence for them. The systems contains two essential components: MGNet for joint semantics encoding of hypothesis and premise, as well as the inference mechanism in multi-granularity by both sentence-level encoding and token-level encoding. Additionally, the system uses a SciFive model to strengthen numerical reasoning for hypotheses that call upon this form of inference.

Vladika and Matthes [68] developed two main models, A pipeline system that dictinctly tackles the two tasks and a combined system capable of learning both task at once by means of sharing representations via applying multi-task learning. Then an ensemble system which merges predictions of those two models produced the final one. The authors discuss formation of the models, their properties and problems encountered, as well providing an analysis of results. The authors' system acquired an F1 Score of 0.798 on the textual entailment task.

Rajamanickam and Rajaraman [69] of the I2R team presented an ensemble approach based on explanations to tackle the challenge of natural language inference on clinical trial data. The proposed approach combines several pre-trained language models by using explanation techniques to guide the selection and weighting of predictions. The authors specifically leveraged feature attributions from the models to identify the most relevant parts of the hypotheses and premises for classification. This information enabled the creation of a more effective ensemble of models compared to each individual model. Using this method, the author attained an F1 Score of 0.701.

Takehana et al. [70] tested two different approaches, first one is based on fine-tuning and ensembling MLMs, and the other is based on prompting LLMs using approaches like Chain-Of-Thought and Contrastive Chain-Of-Thought. For the MLM-based approach, the team fine-tuned multiple MLMs on the NLI4CT training and development data, and then used hard-voting and soft-voting ensemble methods to make predictions on the test set. For the LLM-based approach, the team designed a set of prompts to instruct the models to perform Textual Entailment, using the statement and relevant section in CTR which is the premise.

The experiment led to their best-performing system that achieved an F1 score of 0.662. They also discuss related work, highlighting the recent advancements in Large Language Models and their increasing capabilities in addressing both generative and discriminative tasks.

2.2.1.1 Comparative Analysis

Zhou et. al. [54] who achieved first place in the international SemEval 2023 competition, demonstrating a substantial performance boost in boths of competition. The author proposes a multi-granularity inference system to address the challenges of the A Multi-evidence Natural Language Inference for Clinical Trial Data task on NLI4CT dataset. MGNet for joint semantics encoding of hypothesis and premise, as well as the inference mechanism in multi-granularity by both sentence-level encoding and token-level encoding. Additionally, the system uses a SciFive model to strengthen numerical reasoning task. Author achieved an F1 score of 0.856.

Vladika et. al. [68], introduced an innovative approach to handling Natural Language Inference tasks within clinical trials by developing two primary models. The first model was a pipeline system designed to address tasks sequentially, where each task is solved individually in a step-by-step fashion. This allowed the system to handle distinct parts of the inference process with specialized focus, potentially improving task-specific performance. The second model implemented a combined multi-task learning system, where both tasks were addressed simultaneously.Then

applied an ensemble approach by combining separate models for multiple tasks, then merging their predictions to achieve a final F1 score of 0.798.

Similarly, Rajamanickam et. al. [69] employed an explanation-driven ensemble approach, using feature attributions to guide the ensembling process on the NLI4CT dataset, achieving an F1 score of 0.701. Both studies highlight how ensembling can effectively combine the strengths of different models while mitigating their individual weaknesses. Takehana et al. [70] employed an ensembling system combining fine-tuned Masked Language Models (MLMs) with hard and soft voting to enhance prediction accuracy on the NLI4CT dataset.

2.2.2 Deep Learning Based Systems

Deep learning approaches have become a cornerstone of contemporary research in natural language processing and other domains, as evidenced by their widespread adoption in the existing literature. Numerous studies demonstrate the effectiveness of deep learning models, particularly those based on transformer architectures like BERT and its variants, which have been fine-tuned for tasks such as NLI.

BERT Based Models

Wang et. al. [71] from KnowComp performed Fine-tuned a Pre-trained Language Models for Clinical Trial Entailment Identification. The task involves identifying whether a given claim is entailed by the evidence provided in a set of clinical trial descriptions. Author fine-tuned pre-trained language models like BERT and RoBERTa, on the provided dataset to build a system for this task. The paper details the team's approach, including their model architecture, training procedure, and other technical details. The results show that the KnowComp system achieved strong performance by achieving an F1 Score of 0.764 on the task, demonstrating the effectiveness of their fine-tuning approach.

Volosincu et al. [72] tested if a transformer-based model pretrained on biomedical data could outperform general language models on the task of natural language inference for clinical trial data. The author used an ensemble of biomedical language models, including LinkBERT and other fine-tuned transformer models, to tackle the multi-evidence natural language inference challenge. The models were trained to infer whether a given hypothesis statement is supported, refuted, or neutral based on the provided evidence from clinical trial data. Employing this technique, the author achieved an F1 Score of 0.596. Detailed analysis of the team's performance and the key findings from their approach are presented in the paper.

Alameldin and Williamson [73] of Clemson NLP team's presented a thorough review of their performance as well detailed description of GatorTron which is a transformer-based language model pretrained on clinical texts. The GatorTron model was applied to the task of multi-evidence clinical NLI, where the goal is to determine if a given hypothesis is entailed, contradicted, or neutral based on multiple evidence passages. The paper discusses the model architecture, training, and evaluation on the NLI4CT dataset. Results of this work shows an F1 Score of 0.7065. This work also offers insights into the challenges and performance of the Clemson NLP team's approach to this important task in the clinical natural language processing domain.

Bevan et. al. [74] of MDC team employed fine-tuned transformer models for the textual entailment prediction and evidence retrieval subtasks. For textual entailment, the author customized a RoBERTa-based model using a task-specific dataset. In the case of evidence retrieval, the author fine-tuned a passage retrieval model to locate relevant passages within clinical trial reports.

The results show that their approach achieved competitive performance on both subtasks, demonstrating the effectiveness of fine-tuning transformer models for these natural language inference and information retrieval challenges in the clinical domain. Employing this technique, the author achieved an F1 Score of 0.695.

Feng et. al. [75] presents an NLI approach based on multiple pieces of evidence for clinical trial data, utilizing BioBERT model. The authors aimed to determine the relationship between a hypothesis and a set of evidence items from clinical trials. In this study an NLI model based on BioBERT was developed that integrates multiple pieces of evidence to predict the relationship between a hypothesis and evidence items from clinical trials. The approach achieved strong performance on the NLI4CT datasets for textual enatilament, ranking among the top systems.

The paper reviews related work in the field of deep learning for natural language processing, including studies that have utilized various architectures such as LSTM, Bi-LSTM, and transformer-based models for NLI tasks. The authors also discuss the importance of incorporating external knowledge, such as semantic dependencies, to improve the results of NLI models. The proposed "Role-based Double Roberta-Large" model leverages the strengths of the Roberta-Large architecture and incorporates role-based information to enhance the multi-evidence NLI task.

The researchers demonstrate the effectiveness of their approach through experiments and comparisons with other state-of-the-art models. Overall, this publication contributes to the advancement of NLI techniques, particularly in the context of clinical trial data, and highlights the potential of deep learning-based methods to facilitate the extraction and interpretation of medical evidence, leading to improved personalized patient care [76], which ultimately improves reliablity of medical systems.

LLM Based Models

Liu et. al. [77] developed an interpretable inference system for complex biomedical reasoning tasks. The authors use a Chain-of-Thought (CoT) reasoning methodology, which they further enhance with a self-consistency mechanism. This approach improves the system's accuracy by generating multiple reasoning chains for a given prompt rather than relying on a single chain. Using majority voting, the model

then selects the most consistent result among these multiple chains. The author used GPT-4 to generate rationales for the training and validation sets. By filtering instances not matching labels generated by GPT and labels in dataset and ensuring higher-quality examples for model training. To enhance model efficiency, the authors utilized the Low-Rank Adaptation (LoRA) instruction tuning framework by allowing the model to minimize parameter updates for each task. This method not only improves training efficiency. The authors' self-consistent CoT model demonstrated an F1 Score of 0.80.

Chakraborty [78] explored three main modeling approaches, a discriminative Graph Attention Network (GAT) model utilizing dependency parsing, a generative model based on T5 variants enriched with synthetic data, and a large language model (LLM) approach with GPT-4, both with and without few-shot learning. The LLM-based approach with GPT-4 emerged as the best performer, confirming the effectiveness of larger models in biomedical NLI tasks. Although the team did not use few-shot examples, they suggested that careful prompt engineering and the inclusion of targeted examples could further enhance results. The author achieved an F1 score of 0.76 by using GPT-4.

Zhao et. al. [79] the member of Hw-tsc team explored the capabilities of NLI on large language models, specifically ChatGPT and other pre-trained models, to address this task. They fine-tuned the models on a dataset provided for the task and checked their results on the test set. The results showed that the fine-tuned language models, particularly ChatGPT, were able to achieve strong performance on the task, outperforming other baselines. The authors note that the models were able to capture relevant semantic relationships and infer the relevance of the clinical trial descriptions to the patient conditions. Author has given an extensive summary of the model's performance, including an exploration of their strengths and limitations. The authors also discuss the potential of using large language models for clinical trial matching and the implications for improving patient-trial matching in the future. The author obtained an F1 Score of 0.679.

Pahwa and Pahwa [80] investigated whether fine-tuned cross-encoder language models could outperform the GPT-3.5 model on these phrase inference tasks. They fine-tuned several cross-encoder models and compared them to GPT-3.5 by using evaluation metrics such as F1 score, precision, recall and Accuracy. The results indicate that some fine-tuned cross-encoder models indeed surpassed GPT-3.5 on these phrase inference tasks from clinical trial data. The authors conclude that fine-tuned cross-encoder models can be a promising alternative to large language models for specific tasks involving domain-specific data. This publication thus provides insights into the comparative performance of cross-encoder language models and GPT-3.5 within the specific context of phrase inference from clinical trial data. The author reached an F1 Score of 0.679 through this technique.

Chakraborty [78] who explored three main modeling approaches, a discriminative Graph Attention Network (GAT) model utilizing dependency parsing, a generative model based on T5 variants enriched with synthetic data, and a large language model approach with GPT-4, both with and without few-shot learning. The LLM-based approach with GPT-4 emerged as the best performer, confirming the effectiveness of larger models in biomedical NLI tasks. Although the team did not use few-shot examples, they suggested that careful prompt engineering and the inclusion of targeted examples could further enhance results. The author achieved an F1 score of 0.76 by using GPT-4.

T5 Based Models

Kanakarajan and Sankarasubbu [81] of Saama AI research group opted for an approach based on an instruction-tuned LLM. They evaluated various publicly available LLMs in a zero-shot scenario and fine-tuned Flan-T5 model which performed best for this task. With this system, Author achieved an F1 score of 0.834 on the dataset provided by competition organizers, securing second place on the leaderboard. The authors argue that Large Language Models, due to their remarkable performance in many language processing tasks, are well-suited to tackle this

challenge. They demonstrate that using an instruction-tuned LLM is a promising approach for addressing tasks such as interpreting and searching medical evidence.

Smilga [82] used the Flan-T5 family of models, he experimented with both original and augmented datasets created with GPT-3.5-Turbo. Data augmentation techniques included synonym replacement, syntactic rephrasing, random fact insertion, and meaning reversion, aiming to improve the model's faithfulness (ability to detect semantic changes) and consistency (ability to maintain predictions for semantically similar inputs). However, the author noted a trade-off, as data augmentation often reduced performance on the unaltered dataset, as measured by F1 score. The best-performing model, Flan-T5 XL, was fine-tuned on a combination of original and over 6,000 augmented examples. Author achieved an F1 score of 0.76 using this approach.

Other Models

Guimaraes et. al. [83] used the Mistral-7B model, an open-source LLM, which was adapted to the NLI4CT task through prompt engineering and fine-tuning on a quantized version of the model, combined with a data-augmented training set. Initially, the authors tested several open-source LLMs for their zero-shot and few-shot capabilities, ultimately choosing the Mistral-7B-Instruct-v0.2 model for its ability to handle long input texts, a crucial requirement given that CTRs can exceed 3000 tokens. To maximize the model's performance, they designed a prompt specifically structured for the NLI4CT task. Data augmentation played a significant role in their approach, compensating for the relatively limited dataset. Additionally, the author fine-tuned Mistral-7B using LoRA (Low-Rank Adaptation) in a quantized 4-bit format to support sequences up to 6000 tokens. The author achieved an F1 score of 0.80 by using this approach.

Lee et. al. [84] presented system that integrates several LLMs and performs interventions on input statements to find entailment. The system consists of four main components: data augmentation using ChatGPT, fine-tuning the SOLAR

model with augmented data, using OpenChat for intervention reduction, and label prediction with fine-tuned LLMs. For data augmentation, the team used ChatGPT to rephrase statements in diverse ways, introducing variations in vocabulary, syntax, and numerical formatting to prepare for real-world inference scenarios.

To optimize the model's robustness against manipulated data, the team fine-tuned the SOLAR-10.7B model using instruction tuning and LoRA, incorporating prompts and structured templates to enhance performance on NLI tasks. The author highlights that while the approach was successful, further refinements, including tailored instruction prompts and domain-specific LLMs, could improve performance even further. Author achieved an F1 score of 0.779 using LLM

2.2.2.1 Comparative Analysis

Wang et al. [71] fine-tuned pre-trained language models like BERT and RoBERTa specifically for clinical trial entailment identification. Volosincu et al. [72] expanded on this by employing an ensemble of biomedical language models, including LinkBERT, further highlighting BERT's adaptability to specialized contexts.

Chen et al. [25] utilized BERT for encoding linearized tables and statements in their binary classification of entailment, while integrating techniques like the Latent Program Algorithm (LPA) for program synthesis. Alameldin et al. [73] introduced GatorTron BERT, a transformer-based model pretrained on clinical texts, aimed at multi-evidence clinical NLI tasks. Similarly, Bevan et al. [74] refined transformer models like RoBERTa for textual entailment and evidence retrieval. Feng et al. [75] adopted a NLI approach with the BioBERT model, analyzing relationships between hypotheses and multiple pieces of evidence from clinical trial data.

Alissa et al. [76] presented a "Role-based Double Roberta-Large" model, illustrating the effectiveness of transformer models for navigating the complexities of clinical data. Additionally, researchers like Zhou et al. [54], Vladika et. al. [68], and Rajamanickam et al. [69] employed ensembling methods, combining various deep

learning models to enhance NLI system performance. Collectively, these studies underscore BERT's pervasive influence in advancing research in NLP, particularly within clinical and biomedical contexts.

LLMs, particularly those in the GPT family, have become increasingly prominent in current research, demonstrating capabilities in various natural language processing tasks. For instance, Liu et al. [77] employed a Chain-of-Thought reasoning approach enhanced by a self-consistency mechanism, enabling GPT-4 to generate and validate reasoning chains for complex biomedical reasoning tasks. Similarly, Chakraborty [78] explored multiple methodologies, including a generative model based on T5 variants and an LLM approach with GPT-4, with the latter emerging as the top performer in the NLI task.

Zhao et al. [79] focused on fine-tuning ChatGPT to improve performance in NLI tasks related to clinical trial matching, showcasing the model's adaptability to specific applications. Furthermore, Pahwa et al. [80] investigated the efficacy of cross-encoder language models, comparing their performance against the GPT-3.5 model on phrase inference tasks. Their results revealed that some fine-tuned cross-encoder models could outperform GPT-3.5, yet the continued exploration of LLMs like GPT indicates their significant role in advancing research in clinical and biomedical contexts. This trend emphasizes the versatility and effectiveness of GPT-based approaches across a variety of NLP challenges.

The T5 model, particularly its variants like Flan-T5, has been utilized in recent research, highlighting its effectiveness in various natural language processing tasks. For example, Kanakarajan et al. [81] employed a fine-tuned Flan-T5 model, showcasing its capability as an instruction-tuned large language model that excels in zero-shot scenarios for natural language inference tasks. Kanakarajan et. al. [81] achieved 2nd position in International SemEval 2023 competition. Similarly, Smilga's [82] approach involved a combination of Flan-T5 and GPT-3.5-Turbo, focusing on enhancing model consistency and robustness when detecting semantic

changes. These studies reflect the growing trend of leveraging T5 models to address complex challenges in NLI and related tasks, reinforcing their significance in contemporary deep learning research.

Similarly other large language models, have been employed across some research papers to tackle complex tasks in natural language processing. For instance, Guimaraes et al. [83] fine-tuned the Mistral-7B model using Low-Rank Adaptation (LoRA) on a quantized, data-augmented dataset, enabling it to effectively manage long text inputs—an essential requirement for processing extensive clinical trial data. Similarly, Lee et al. [84] explored multiple LLMs for entailment detection, specifically fine-tuning the SOLAR-10.7B model through instruction tuning and LoRA. They also incorporated data augmentation techniques via ChatGPT, which significantly enhanced the model's robustness and performance on NLI tasks. These examples illustrate the versatility and effectiveness of various Deep Learning models in advancing research within the field.

2.2.3 Machine Learning Based Systems

Recent literature reveals that relying solely on a single machine learning approach is becoming less common, especially in complex tasks such as NLI. However, Hybrid approaches are used which make use of Machine learning along with some other technique.

2.2.4 Rule Based Systems

Raabia et. al. [38] developed CustNER, a rule-based named-entity recognition system that has improved recall compared to other NER models. CustNER uses a set of 7 rules to identify named entities in text. The rules are designed to improve recall by capturing entities that may be missed by machine learning-based NER models. Evaluation results show that CustNER has higher recall compared to other rule-based and machine learning-based NER systems, while

maintaining a high level of precision. CustNER was able to identify more named entities, particularly rare or uncommon ones, that were missed by other NER models. CustNER can be a useful tool for applications that require high recall in named entity recognition, such as information extraction, question answering, and knowledge base population. The rule-based approach also makes CustNER more interpretable and easier to customize for specific domains compared to black-box machine learning models.

2.2.5 Hybrid Approach

Mohamed et. al. [85] employed a hybrid method that combined rule-based strategies along with Machine Learning i.e. TF-IDF and RBF Kernel similarity technique, to predict labels by assessing text similarity and relevance. Their approach achieved an F1 score of 0.667 which is low as compared to deep learning approaches but rule based system operates with limited resources and works better where high precision is paramount or available data is limited.

2.2.6 Other Approaches

Among deep learning approaches, Chakraborty [78] explored a discriminative Graph Attention Network (GAT) model utilizing dependency parsing but its perfomrance is satisfactory. Neves [86] introduced the Bf3R model, an ontology-based text similarity approach designed for textual entailment and evidence retrieval on the Clinical Trial Dataset i.e. NLI4CT.

Chen et. al. [87] introduced TabFact, an extensive new dataset designed for fact verification using table-based information. This dataset was specifically created to explore the challenge of entailment, which involves logical deduction, within the context of semi-structured data presented in tabular form. The TabFact dataset features over 16,000 Wikipedia tables and more than 118,000 statements that are annotated as either entailment or contradiction in relation to the information

contained within these tables. This method helped the author to reach an F1 Score of 0.709 on textual entailment task.

Corrêa Dias et al. [88] used the approach of supervised contrastive learning at the pair level. They trained the EvidenceSCL system on two datasets created from NLI4CT and used some data from other datasets related to NLI. The study demonstrates that the proposed approach effectively addresses both objectives of NLI4CT. This work presents an innovative approach utilizing supervised contrastive learning to tackle the task of pair-level sentence classification and retrieval of evidence within the NLI4CT framework. The author obtained an F1 Score of 0.666. While the results are promising, further enhancements are needed to achieve better performance.

Conceição et al. [89] explored how incorporating domain-specific ontologies can significantly boost the accuracy of natural language inference models, which are crucial for tasks like text understanding and reasoning. The best-performing system from the experiment achieved an F1 score of 0.661. The work is based on natural language inference and the use of domain knowledge to enhance such systems.

2.2.6.1 Comparative Analysis

Chakraborty [78] investigated multiple DL approaches as well as GAT. Neves [86] used an ontology-based text similarity approach. Whereas, Chen et. al. [87] an extensive new dataset i.e. TabFact, designed for fact verification using table-based information. All such approaches have been seen less in recent literature and not achieved high score in terms of F1 Score.

Interestingly, while these innovative approaches show promise, they have been underrepresented in recent literature. The lack of high-performance outcomes suggests that further investigation and refinement are needed to enhance the efficacy of these methods. Moreover, there is an opportunity for developing new benchmarks that incorporate diverse datasets.

TABLE 2.2: Top Performer on Textual Entailment for NLI4CT Dataset

Sr. No.	Reference	Technique	F1 Score
1	Zhou et al., 2023 [54]	DL–Transformer + LSTM	85.60%
2	Kanakarajan et al., 2023 [81]	DL–Flan T5 (LLM)	83.40%
3	Liu et al., 2024 [77]	DL–GPT-4 + Instruction-tune an LLM	80.00%
4	Guimarães et al., 2024 [83]	DL–Prompt learning on Mistral-7B LLM	80.00%
5	Vladika et al., 2023 [68]	DL–BERT + Multi-Task Learning (MTL)	79.80%
6	Lee et al., 2024 [84]	DL–ChatGPT v3.5 + fine-tuned SOLAR model	77.90%
7	Wang et al., 2023 [71]	DL–Fine tuning of DeBERTa-v3-large	76.40%
8	Chakraborty, 2024 [78]	DL–GPT-4 + Graph Attention Network	76.00%
9	Smilga, 2024 [82]	DL–Flan-T5 + Data Augmentation	76.00%
10	Chen et al., 2019 [87]	DL–BERT and TabFact Dataset	70.90%
11	Alameldin et al., 2023 [73]	DL–GatorTron BERT	70.50%
12	Rajamanickam et al., 2023 [69]	DL–Text to Text Transfer Transformer (T5)	70.10%
13	Bevan et al., 2023 [74]	DL–Fine-tuned BioLinkBERT	69.50%
14	Zhao et al., 2023 [79]	DL–Prompt learning based on ChatGPT	67.90%
15	Pahwa et al., 2023 [80]	DL–Fine-tuned BioLinkBERT	67.90%
16	Feng et al., 2023 [75]	DL–Fine-tuned BioBERT	67.90%
17	Alissa et al., 2023 [76]	DL–Role-based Double Roberta-Large	67.00%
18	Noor Mohamed et al., 2023 [85]	Rule-based + TFIDF + RBF Kernel similarity	66.70%

2.2.7 Analysis of Failure Cases

Upon analyzing Training Dataset which contains 500 Adverse Event related cases, we have found that in Clinical Trial statements, some terms occur frequently as shown in Table 2.3. One such example is Adverse Event, which represent disease or condition. Adverse Event keyword / Adverse Event Name and Primary Trial / Secondary Trial is used in almost all of the statements. Although cohort occurrence percentage is 43.23% but if cohort is not mentioned it represents available cohorts (either cohort 1 or both cohorts).

TABLE 2.3: Occurrence of Terms in CTR (AEs Section)

Term	**Occurrence in Statements**
Adverse Event Keyword / Name	100.00%
Primary Trial / Secondary Trial	99.60%
Record / Suffered / Observed ...	61.09%
Cohort 1 / Cohort 2 / Cohorts	43.23%

System of Zhou et. al. [54] failed on eight cases which are grouped into 5 types based on similarity. Analyzing failure cases thoroughly, it can be seen from Table 2.5 that each statement can be decomposed into fixed terms and these terms occur again and again in CTR statements, summary of which is depicted in Table 2.3. The frequency of these recurring terms shows a clear pattern.

Furthermore, statements contain similar terms / synonyms which can be replaced with common terms e.g. similar terms in input statements such as suffered, observed, recorded etc. can be replaced with standard keyword i.e. Record. These standard keywords can be stored in keywords dictionary which aims to map similar terms with standard keyword. Since statement can be decomposed into limited terms which can be further simplified by replacing similar terms with single term or keyword this makes it easier to design rules. This is because structure of English grammar usually remain same in contemporary system. English grammar follows well-defined rules for sentence construction, word order, and phrase structure, which are widely understood and consistently applied across various NLP models

and applications. The predictability of English grammar also aids in the development of rule-based and machine learning systems, as models can rely on these grammatical patterns to improve accuracy and interpret meaning with minimal ambiguity. This inherent regularity in the language structure, therefore, serves as a foundation upon which NLP techniques can build more reliable, adaptable, and precise solutions.

TABLE 2.4: Summary of Adverse Event Cases and Types

Sr No	**Case Example**	**Type**
1	less than 5% of patients in the primary trial suffered AEs	Type 1
2	less than 10 patients in the primary trial suffered AEs	Type 1
3	the secondary trial recorded slightly more total adverse events in its patient cohorts than the primary trial.	Type 2
4	the primary trial and the secondary trial recorded the same total number of adverse events in their patient cohorts.	Type 3
5	All Infections and Fever cases in the primary trial were for patients in cohort 1.	Type 4
6	All Infections and Infestations cases in the primary trial were for patients in cohort 1.	Type 4
7	The only types of AEs observed in patients from the secondary trial were Eyelid oedema, Upper gastrointestinal haemorrhage, and Chest pain; no AEs were recorded in the primary trial.	Type 5
8	The only types of AEs observed in patients from the secondary trial were Eyelid oedema and Chest pain; no AEs were recorded in the primary trial.	Type 5

Considering statements at Sr No 1 and 2 in Table 2.5, these cases are similar in nature and can be decomposed into following parts after conversion to standard terms.

1. <5% or <10
2. primary trial
3. record
4. aes

Terms 'primary trial', 'record' and 'aes' can be compared with standard terms using string matching method, whereas, expressions (<n or <n %,) can be compared using regex. These standard terms and regex(s) are building blocks of a Rule. Comparing a terms in input statement with terms defined in a rule is known as Rule Matching, which decides whether the rule processing will be applied on input statement or not. Consider three examples of dataset.

Rule Matching refers to the process of comparing terms in an input statement with those defined in a rule. This comparison is crucial as it determines whether the specified rule processing will be applied to the input statement. If the terms match, the corresponding rules can be activated to process the input appropriately.

To illustrate this concept, let's consider three examples from a dataset that highlight the application of these rules in real-world scenarios. These examples will demonstrate how effective rule matching can facilitate better data processing and enhance the accuracy of outputs.

Original Statement

- AEs were not recorded for the primary trial or the secondary trial
- Less than 5% of patients in the primary trial suffered AEs
- Less than 10 patients in the primary trial suffered AEs

After removing stopwords and tagging keywords, we can get these encoded statements.

- `[[aes]] [[exp:=0]] [[record]] [[ptost]]`
- `[[exp:<5%]] patients [[pt]] [[record]] [[aes]]`
- `[[exp:<10]] patients [[pt]] [[record]] [[aes]]`

From above list it is evident that all of the above encoded statements can be matched by matching 4 tags

Table 2.5: Analysis of Failure Cases

No.	Statement	Type	AEs	Location	Keyword 1	Expression	Keyword 2	Cohort	Comparison With
1	less than 5% of patients in the primary trial suffered AEs	Single		primary trial	suffered	less than 5%	AEs		
2	less than 10 patients in the primary trial suffered AEs	Single		primary trial	suffered	less than 10	AEs		
3	the secondary trial recorded slightly more total adverse events in its patient cohorts than the primary trial.	Comparison		Secondary Trial	recorded	slightly more	adverse events	cohorts	Primary Trial
4	the primary trial and the secondary trial recorded the same total number of adverse events in their patient cohorts.	Comparison		primary trial	recorded	same total	adverse events	cohorts	Secondary Trial
5	All Infections and Fever cases in the primary trial were for patients in cohort 1.	Single	Infections, Fever	primary trial	were for			cohort 1	
6	All Infections and Infestations cases in the primary trial were for patients in cohort 1.	Single	Infections, Infestations	primary trial	were for			cohort 1	
7a	The only types of AEs observed in patients from the secondary trial were Eyelid oedema, Upper gastrointestinal haemorrhage and Chest pain	Comparison	Eyelid oedema, Upper gastrointestinal haemorrhage and Chest pain	Secondary Trial	Observed		AEs		
7b	no AEs were recorded in the primary trial	Comparison		primary trial	recorded		no AEs		
8a	The only types of AEs observed in patients from the secondary trial were Eyelid oedema and Chest pain	Comparison	Eyelid oedema and Chest pain	Secondary Trial	Observed		AEs		
8b	no AEs were recorded in the primary trial	Comparison		primary trial	recorded		no AEs		

2.2.8 Identified Research Gaps

Literature reviews indicate that Deep Learning models are increasingly employed due to their superior performance and flexibility. While deep learning systems excel in processing vast amounts of data and identifying complex patterns, they can sometimes miss nuanced or specific instances due to limitations in their training data or inherent biases in their learning algorithms.Leading Deep learning based NLI system [54] with remarkable performance also has almost 15% improvement margin which can be enhanced by employing a hybrid system.

Rule-based NLP systems excel in specific domains. These systems exhibit high accuracy and precision, operating cause-and-effect principles. They require minimal data to make decision. Furthermore, Rule-based NLP systems can make informed decisions quickly and efficiently. Wang et. al. [67] highlight the risk for deep learning models to incorporate malicious information amplified by generative AI integration into social media, which is not the case in rule-based techniques, as rules are made and verified by human experts.

After a thorough analysis of the failure cases observed in Zhou et al.'s system, we identified a pattern where certain words frequently recur in the hypothesis statements. The identification of these recurrent vocabulary patterns presents a significant opportunity to enhance the system's performance. By leveraging these insights, we can develop targeted strategies to address the specific challenges. Furthermore, for the specific failure cases identified during our analysis, designing targeted rules based on the recurrent terms in the hypothesis can provide a more structured approach to addressing these challenges.

A rule-based system serves as a valuable complement to deep learning models by effectively addressing cases that may be overlooked by the latter. By integrating rule-based system with deep learning system, we can create a hybrid approach that not only boosts overall performance but also ensures more reliable outputs for challenging scenarios.

Chapter 3

Proposed System - A Hybrid Approach

3.1 Introduction

In this chapter, a practical exploration of the proposed system model and the detailed design of a hybrid system is undertaken. Chapter 3 serves as a direct continuation of the comprehensive study initiated in Chapter 1, where the motivation and overarching objectives of the research were introduced.

Rather than starting from scratch, we utilized leading Deep Learning System [54] along with Rule-based System for cases missed by deep learning system. We carefully analyzed the patterns of deep learning system failure cases and developed rules to identify each type of missed pattern. The rule development is an iterative process where rules are tested and refined if results are not satisfactory. The proposed system incorporates a combination of fixed and generic rules, designed to address multiple cases effectively. The proposed hybrid system effectively combines the strengths of both Deep Learning and Rule-based System. This system finds inference using Deep Learning system and then passes designed to address multiple cases and enhance overall accuracy. This integration allows for improved performance in complex scenarios.

3.2 Proposed Hybrid System

Proposed system is a combination of Deep Learning system and Rule-based system as shown in Figure 3.3. First NLI4CT dataset and CTR files are passed to Deep Learning system for finding inference. Result of this system is passed to Rule-based system, which apply rules designed for cases where deep learning system filed. Rule-based system overwrites output file of Deep learning system to find final inference results.

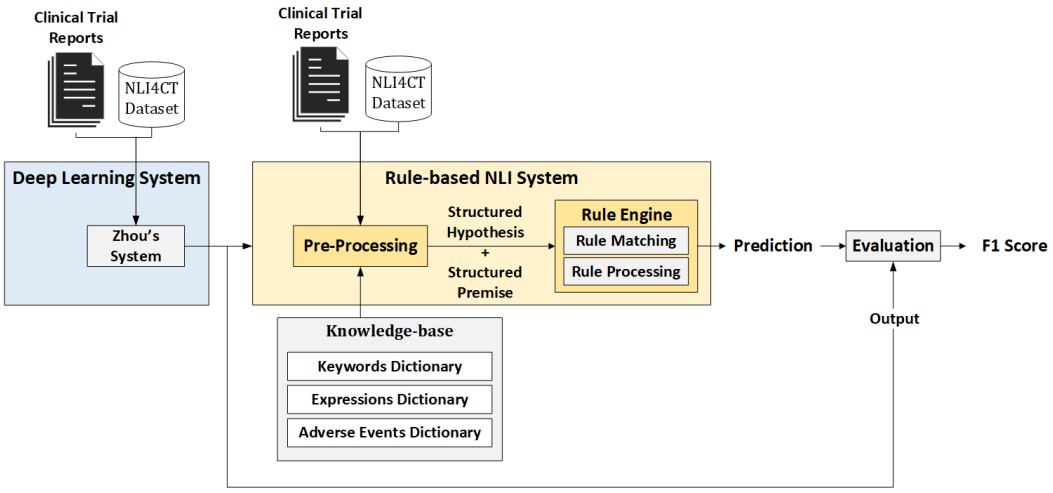


FIGURE 3.1: Block Diagram of Proposed Hybrid System

3.2.1 Deep Learning System (Zhou's System)

Zhou et al. attained 1st position in International SemEval 2023 Competition. He addressed the NLI4CT task, which involves determining whether hypotheses are supported by CTRs and retrieving the relevant evidence. The authors propose a multi-granularity system called the MGNet to tackle both textual entailment and evidence retrieval as demonstrated in Figure 3.2. Additionally, the system enhances its numerical inference capabilities by utilizing the T5-based model, SciFive, which is pre-trained on medical texts. To further improve prediction stability and consistency, the authors employ model ensembling and joint inference methods. The performance of this approach is notable, achieving F1-scores of 0.856 for

textual entailment and 0.853 for evidence retrieval, thereby outperforming existing models in these areas.

3.2.1.1 Model Description

3.2.1.2 Multi-granularity Inference Network:

- **Joint Semantics Encoder:** Uses a transformer-based model to learn the contextual representation of hypotheses and premises, formatted as a sequence.
- **Sentence-level Encoder:** Processes the pooled token-level representations of sentences using two approaches: BiLSTM and a transformer encoder, to extract contextual semantics.
- **Token-level Encoder:** Provides fine-grained representations for individual sentences, aiding evidence retrieval. Implemented through either a BiLSTM or max-pooling layer.
- **Classifiers:** Implemented with simple structures for both tasks, utilizing MLPs to determine the probability of textual entailment and evidence support.

SciFive Model:

Addresses the system's weaknesses in numerical inference by predicting probabilities for "entailment" and "contradiction" based on the input sequence.

Joint Inference Network:

Enhances consistency in predictions by evaluating pairs of hypotheses that share the same premise, employing a transformer encoder to analyze their relationships.

Soft Ensembling:

Utilizes a cross-validation approach to summarize inference results from various models, improving the overall performance by averaging predictions.

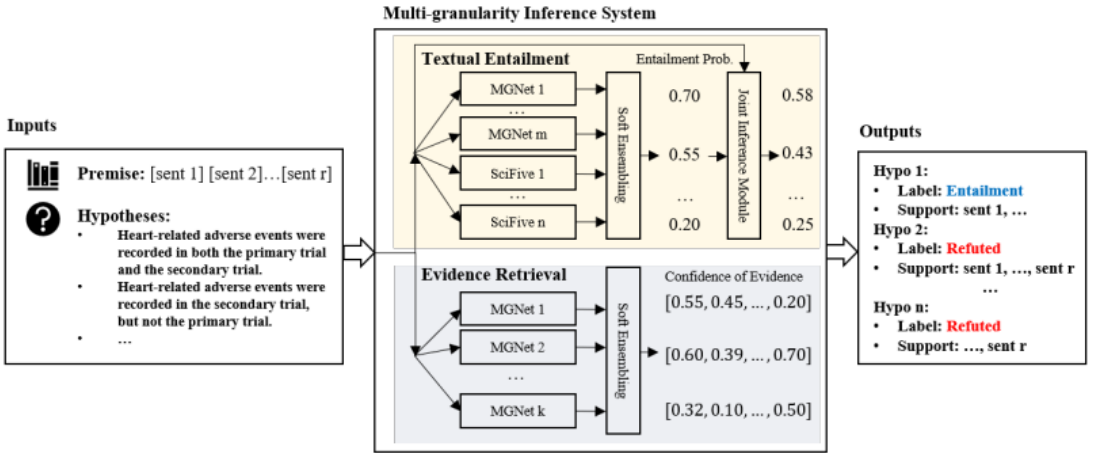


FIGURE 3.2: An Overview of Proposed Multi Granularity System. Figure from [54]

The proposed multi-granularity system effectively uses ensembling system to tackle the complexities of textual entailment and evidence retrieval in clinical contexts. The system demonstrates superior performance in comparison to existing methods. However, despite its strong results, Zhou's system still failed in many cases, indicating areas for improvement. Designing specific rules and implementing a rule-based system could enhance its reliability and accuracy in challenging scenarios.

3.2.2 Rule Based System

In this section, we delve into the core of the proposed method for attaining high prediction accuracy with selected statements. The rule-based approach emphasizes understanding the structure of statements, formulating rules, and subsequently fine-tuning them to enhance performance. The primary focus is on pre-processing and developing more generalized rules.

Our system does not address the NLI problem for all possible cases. Instead, it targets cases that the Zhou's system has missed by applying specific rules designed to identify these overlooked cases. Following this, a rule engine processes the data to determine whether the premise and hypothesis entail or contradict each other. Complete process of designing rule based system is shown in Figure 3.3.

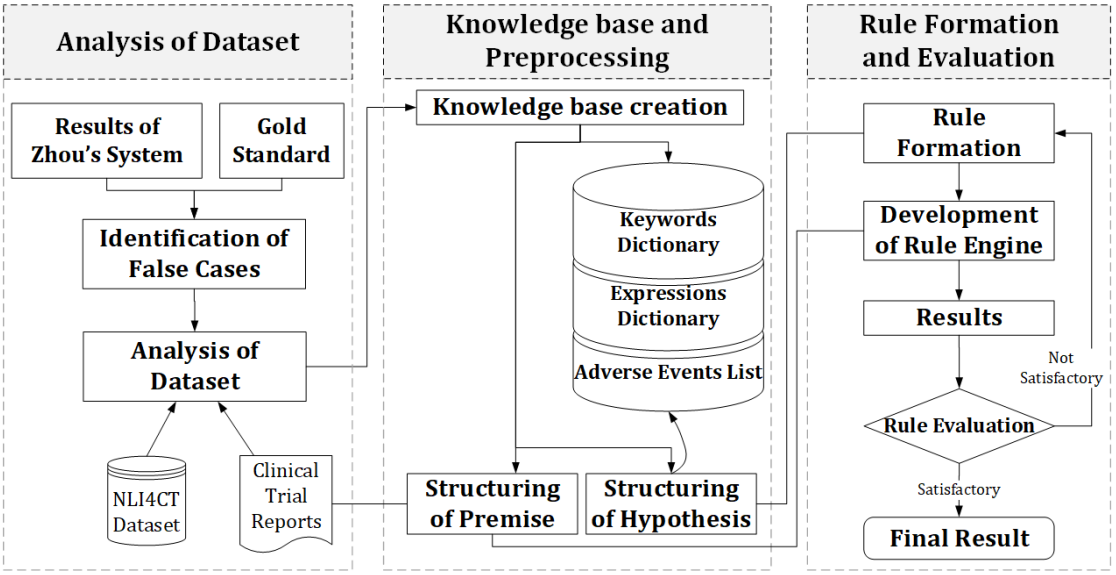


FIGURE 3.3: Flow of Work

Upon comparing the output of Zhou's system with the Gold Standard, we identified several false cases where the system's predictions diverged from the expected results. These false cases highlight specific areas where the system may be underperforming, suggesting that certain patterns or nuances in the data are not being fully captured. A detailed analysis of these discrepancies is essential to pinpoint the underlying causes, which could range from limitations in model training data to insufficient handling of unique medical terminology or complex sentence structures. By thoroughly examining these false cases, we can identify potential enhancements, refine the system's rule set or model architecture, and ultimately improve the accuracy and reliability of the existing system in alignment with the Gold Standard.

3.2.2.1 Analysis of Dataset

The analysis of the dataset revealed consistent and significant patterns manifested as recurring phrases, numerical expressions, and specific medical terminology. These patterns are instrumental in illuminating key language structures and common medical descriptors, which are vital for effective communication in the medical field. By systematically examining these elements, we can accurately

capture essential information that underpins clinical trial data and related documentation. Furthermore, the identification of these consistent elements not only aids in understanding the language of medical literature but also provides a robust foundation for developing precise, targeted rules.

These rules can serve as a framework for enhancing overall performance in various natural language processing tasks, particularly in the realm of clinical text analysis. By leveraging these insights, we can improve the accuracy of models designed to interpret complex clinical data

3.2.2.2 Knowledge Base Creation

In Knowledge Base creation process, we created a comprehensive keywords dictionary to tag recurring keywords efficiently, ensuring consistent recognition of terms that frequently appear across documents. Additionally, an expression dictionary was developed to capture and tag complex expressions, such as numerical and medical phrases, which aids in standardizing diverse forms of expression within the text. To address adverse events specifically, we compiled an adverse events list to tag and identify these events accurately within the dataset. This structured approach to tagging forms a solid foundation for downstream processing, supporting higher precision and relevance in identifying critical data elements in the dataset.

3.2.2.3 Structuring of Premise

In the Preprocessing stage, structuring the premise is essential to organize data effectively for downstream analysis. The statements are saved in a test.json file, while the CTRs are stored in a folder, with each CTR kept in a separate file.

In this step, we extract specific information from the Adverse Events section of CTRs and populate a custom data structure to capture key details, such as the type of adverse events, affected patient counts, and other relevant metrics. CTR Data structure is illustrated in the table 3.1.

We employ advanced natural language processing techniques to meticulously extract specific information from the Adverse Events section of CTRs. This structured representation enables us to effectively query and analyze the data, leading to more accurate and insightful conclusions.

TABLE 3.1: Adverse Events Data Structure Template

CTR Data Structure

```
mapitem = {
    'id': '',
    'statement': '',
    'primaryId': '',
    'secondaryId': '',
    'primary_evidence': {
        'Cohort1': {
            'Total': 0,
            'TotalPatients': 0,
            'Percentage': 0,
            'Adverse Events': [
                # Adverse events here
            ]
        },
        'Cohort2': {
            'Total': 0,
            'TotalPatients': 0,
            'Percentage': 0,
            'Adverse Events': [
                # Adverse events here
            ]
        }
    },
    'secondary_evidence': {
        'Cohort1': {
            'Total': 0,
            'TotalPatients': 0,
            'Percentage': 0,
            'Adverse Events': [
                # Adverse events here
            ]
        },
        'Cohort2': {
            'Total': 0,
            'TotalPatients': 0,
            'Percentage': 0,
            'Adverse Events': [
                # Adverse events here
            ]
        }
    }
}
```

3.2.2.4 Structuring of Hypothesis

The preprocessing step for structuring the hypothesis transforms raw input statements into a refined format that enhances analytical clarity. This structured hypothesis is created through a series of systematic steps which are as under:

1. **Convert Words to Numbers**: This step involves replacing written words with their numeric equivalents, such as changing "one" to "1".

2. **Identify and Replace Keywords**: Similar words are substituted with predefined keywords, for example, replacing "primary trial" with "[[pt]]".

3. **Convert Expressions**: This step entails translating textual expressions into numeric format, such as converting "more than half" to "[[exp:>50%]]".

4. **Eliminate Stopwords**: Stopwords are removed, with the exception of those necessary for rule processing.

5. **Tag Adverse Events**: In this step, adverse events are tagged, such as tagging "Fever" as "<<FEVER >>".

Considering above mentioned steps, the table 3.2 below demonstrates the transformation of an input statement through each step.

3.2.2.5 Rule Formation

The rule formation process, as demonstrated in the table 3.3, shows how original statements are transformed into structured expressions. Each of these four statements is systematically processed to create a structured hypothesis containing exactly four tags: an expression for quantity or percentage ([[exp]]), a patient or cohort reference ([[pt]], [[st]], or [[c1pt]]), an indicator of observation or record ([[record]]), and the type of event ([[aes]] for adverse events).

TABLE 3.2: Transformation of Input Statements

Step	**Processed Statement**
Input Statement	Across both the primary trial and the secondary trial over ten deaths were recorded in the adverse events.
Convert Words to Numbers	Across both the primary trial and the secondary trial over 10 deaths were recorded in the adverse events.
Identify and Replace Keywords	Across [[ptnst]] over 10 deaths were [[record]] in the [[aes]]
Convert Expressions	Across [[ptnst]] [[exp:>10]] deaths were [[record]] in the [[aes]]
Eliminate Stopwords	Across [[ptnst]] [[exp:>10]] deaths [[record]] [[aes]]
Mark Adverse Events	Across [[ptnst]] [[exp:>10]] <<DEATH >>[[record]] [[aes]]

TABLE 3.3: Analysis of Structured Hypothesis for Rule Formation

Original Hypothesis	**Structured Hypothesis**
Less than 5% of patients in the primary trial suffered AEs	[[exp: <5%]] patients [[pt]] [[record]] [[aes]]
5 patients in the secondary trial observed adverse events	[[exp: =5]] patients [[st]] [[record]] [[aes]]
More than half of the patients in cohort 1 of the primary trial recorded AEs	[[exp: >50%]] patients [[c1pt]] [[record]] [[aes]]
None of the patients in the primary trial experienced AEs	[[exp: = 0]] patients [[pt]] [[record]] [[aes]]

3.2.2.6 Rule Engine

The main steps in a rule engine are pattern matching and rule processing:

- **Rule Matching**: In this step, the rule engine compares facts and data against defined rules to identify which rules are applicable. Pseudo-code for Rule Matching is given as Algorithm 1.

- **Rule Processing**: After identifying the relevant rules through pattern matching, the rule engine executes the actions or consequences defined in those rules [90]. This is the phase where the rules are actually applied to modify the system state or generate output.Rule Matching Pseudo-code is given as Algorithm 2.

The separation between pattern matching and rule processing is a fundamental design principle in rule engine architectures.

Let S represent the statement, $R[]$ denote the array of regex rules, and D be the adverse events dictionary. A statement S qualifies for Rule Processing if it matches all regex in $R[]$. Once a statement is deemed qualified, the dictionary D is then traversed to verify whether the data it contains meets the conditions specified by the rule. Further details on the rules are discussed below.

3.2.3 Developed Rules

Rule 1

Rule 1 is a composite rule capable of incorporating functions of group of rules. It identifies keywords and expressions within a hypothesis. Hypothesis and text obtained after pre-processing are given below

Hypothesis: less than 5% of patients in the primary trial suffered AEs.
Processed Text: [[exp:<5%]] patients [[pt]] [[record]] [[aes]]

Where, possible values of terms enclosed in round brackets () in Figure 3.4 are given below:

- **Operator:** `>, <, and =`
- **Number:** Any decimal number

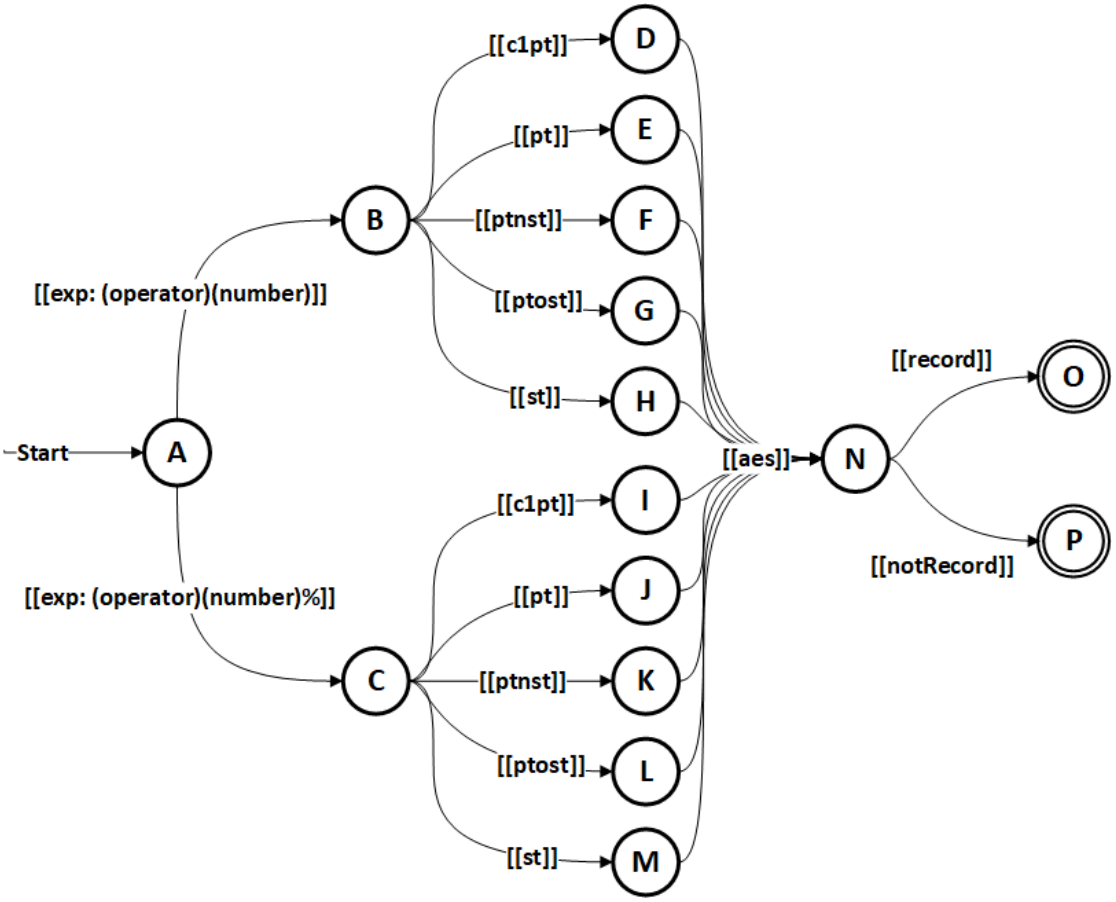


FIGURE 3.4: Automata Representation for Rule 1

In order for this rule to match a statement, it needs to match 4 keywords as mentioned below.

- `[[expression]]`
- `[[c1pt]]` or `[[pt]]` or `[[ptnst]]` or `[[ptost]]` or `[[st]]`
- `[[aes]]`
- `[[record]]` or `[[notrecord]]`

Automata representation of this rule is shown in Figure 3.4. After rule matching, rule processing will check whether the hypothesis entails the CTR Premise. The CTR Premise for this hypothesis is given below. This entails evaluating the logical structure of the hypothesis against the conditions specified in the premise, ensuring that all necessary logical relationships are satisfied.

CTR Premise


```
Primary Evidence:
{'Cohort1': {'Adverse Events': [{'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Gastroesophageal
                                 reflux disease *',
                                 'Patients': '1',
                                 'Percentage': '2.86',
                                 'Synonyms': '',
                                 'Total Patients': '35'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Ductal carcinoma
                                 in situ *',
                                 'Patients': '1',
                                 'Percentage': '2.86',
                                 'Synonyms': '',
                                 'Total Patients': '35'}],
             'Percentage': 5.71,
             'Total': 2,
             'TotalPatients': 35},
 'Cohort2': {'Adverse Events': [],
             'Percentage': 0,
             'Total': 0,
             'TotalPatients': 0}}
Secondary Evidence:
None
```


Hypothesis states adverse events in primary trial should be less than 5% however it is obvious from CTR Premise that Pirmary trial has total adverse events 5.71 which is greater than 5%. Thus, contradiction occurs between hypothesis and CTR premise, which is what Gold standard declares.

Rule 2

Rule 2 is a relatively straightforward rule, designed to match specific keywords within a hypothesis.The simplicity of this rule lies in its ability to identify and extract relevant terms that are pivotal for understanding the hypothesis' meaning and context. Hypothesis along with its pre-processed version are given below

Hypothesis: the secondary trial recorded 11 more Aes than the primary trial.

Processed Text: [[st]] [[record]] [[num:11]] [[moreAesThan]] [[pt]]

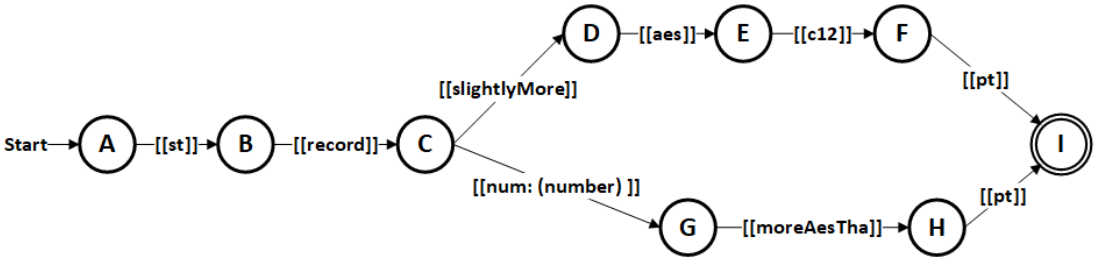


FIGURE 3.5: Automata Representation for Rule 2

Where, (number) in Figure 3.5 can be Any decimal number.

In order for this rule to match a statement, it needs to match keywords in one of the two expression groups, which are mentioned below.

Expressions Group 1

- `[[st]]`
- `[[record]]`
- `[[slightlymore]]`
- `[[aes]]`
- `[[c12]]`
- `[[pt]]`

Expressions Group 2

- `[[st]]`
- `[[record]]`
- `[[number]]`

- `[[moreAesThan]]`
- `[[pt]]`

Automata representation of this rule is shown in Figure 3.5. After rule matching, rule processing will check whether the hypothesis entails the CTR Premise. The CTR Premise for this hypothesis is given below.

CTR Premise

```
Primary Evidence:
{'Cohort1': {'Adverse Events': [],
             'Percentage': 0.0,
             'Total': 0,
             'TotalPatients': 17},
 'Cohort2': {'Adverse Events': [],
             'Percentage': 0,
             'Total': 0,
             'TotalPatients': 0}}
Secondary Evidence:
{'Cohort1': {'Adverse Events': [{'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Febrile neutropenia *',
                                 'Patients': '4',
                                 'Percentage': '16.00',
                                 'Synonyms': '',
                                 'Total Patients': '25'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Neutropenia *',
                                 'Patients': '3',
                                 'Percentage': '12.00',
                                 'Synonyms': '',
                                 'Total Patients': '25'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Diarrhoea *',
                                 'Patients': '1',
                                 'Percentage': '4.00',
                                 'Synonyms': '',
                                 'Total Patients': '25'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Intestinal
                                 perforation *',
                                 'Patients': '1',
                                 'Percentage': '4.00',
                                 'Synonyms': '',
```

```
                                'Total Patients': '25'},
                               {'Abbreviation': '',
                                'Hypernyms': '',
                                'Name': 'Rectal haemorrhage *',
                                'Patients': '1',
                                'Percentage': '4.00',
                                'Synonyms': '',
                                'Total Patients': '25'},
                               {'Abbreviation': '',
                                'Hypernyms': '',
                                'Name': 'Stomatitis *',
                                'Patients': '1',
                                'Percentage': '4.00',
                                'Synonyms': '',
                                'Total Patients': '25'},
                               {'Abbreviation': '',
                                'Hypernyms': '',
                                'Name': 'Vomiting *',
                                'Patients': '1',
                                'Percentage': '4.00',
                                'Synonyms': '',
                                'Total Patients': '25'},
                               {'Abbreviation': '',
                                'Hypernyms': '',
                                'Name': 'Fatigue *',
                                'Patients': '1',
                                'Percentage': '4.00',
                                'Synonyms': '',
                                'Total Patients': '25'},
                               {'Abbreviation': '',
                                'Hypernyms': '',
                                'Name': 'Multi-organ failure *',
                                'Patients': '1',
                                'Percentage': '4.00',
                                'Synonyms': '',
                                'Total Patients': '25'},
                               {'Abbreviation': '',
                                'Hypernyms': '',
                                'Name': 'Erysipelas *',
                                'Patients': '1',
                                'Percentage': '4.00',
                                'Synonyms': '',
                                'Total Patients': '25'},
                               {'Abbreviation': '',
                                'Hypernyms': '',
                                'Name': 'Pseudomembranous
                                colitis *',
                                'Patients': '1',
                                'Percentage': '4.00',
                                'Synonyms': '',
                                'Total Patients': '25'}],
                'Percentage': 44.0,
```

```
            'Total': 11,
            'TotalPatients': 25},
 'Cohort2': {'Adverse Events': [],
            'Percentage': 0,
            'Total': 0,
            'TotalPatients': 0}}
```

The hypothesis correctly predicts that the secondary trial has 11 more adverse events than the primary trial. This is evident from the CTR premise, which states that the secondary trial has 11 adverse events, while the primary trial has none. Consequently, both the proposed system and the gold standard correctly identify this entailment relationship.

Rule 3

Below is one of the hypothesis (input statements) that this rule can match, along with the pre-processed text.

> **Hypothesis:** The primary trial and the secondary trial recorded the same number of adverse events in their cohorts.
> **Processed Text:** [[ptnst]] [[record]] [[equalnum]] [[aes]] [[c12]]

In order for this rule to match a statement, it needs to match 5 keywords as mentioned below. Due to absence of expressions and adverse events, this rule considered one of the simpler rules.

- `[[pt]]` or `[[ptnst]]`
- `[[record]]`
- `[[equalnum]]` or `[[equaltotalnum]]`
- `[[aes]]`
- `[[c12]]`

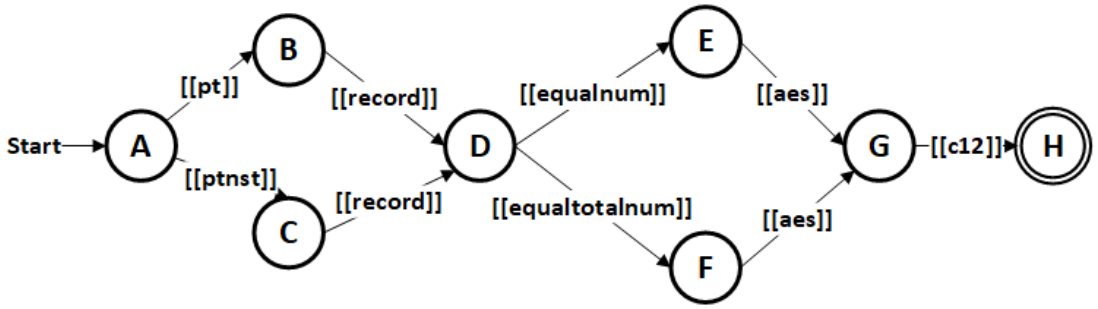


FIGURE 3.6: Automata Representation for Rule 3

Automata representation of this rule is shown in Figure 3.6.

Upon successful rule matching, the system transitions to the rule processing phase. During this phase, the system rigorously assesses the logical relationship between the hypothesis and the CTR Premise. The primary objective of this analysis is to ascertain whether the hypothesis can be logically inferred from the information provided within the Clinical Trial Report (CTR) Premise.

This process involves a careful examination of the relationships between the premise and the hypothesis, focusing on the key elements and statements contained in the CTR that may support or contradict the claims made in the hypothesis.

CTR Premise

```
Primary Evidence:
{'Cohort1': {'Adverse Events': [],
             'Percentage': 0.0,
             'Total': 0,
             'TotalPatients': 34},
 'Cohort2': {'Adverse Events': [],
             'Percentage': 0,
             'Total': 0,
             'TotalPatients': 0}}
Secondary Evidence:
{'Cohort1': {'Adverse Events': [],
             'Percentage': 0.0,
             'Total': 0,
             'TotalPatients': 27},
 'Cohort2': {'Adverse Events': [],
             'Percentage': 0.0,
             'Total': 0,
             'TotalPatients': 27}}
```

Since Total value in cohot 1& 2 of Primary Trial and Total value in cohot 1& 2 of Secondary Trial are equal, this indicates hypothesis entails the CTR premise. Gold standard shows contradiction as result and so the rule engine of proposed system.

Rule 4

Rule 4 is categorized as a composite rule, as it integrates the functionalities of multiple individual rules to enhance its matching capabilities. By combining various criteria for analysis, this rule significantly broadens the spectrum of hypotheses it can accommodate, making it a versatile tool in the inference process. Specifically, Rule 4 operates by matching not only expressions and adverse event names but also a curated selection of keywords that are pivotal in the context of clinical trials. This rule involves matching expressions, adverse event names and keywords. Hypothesis along with its pre-processed text are given below

Hypothesis: Across both the primary trial and the secondary trial only one death was recorded in the adverse events.

Processed Text: [[ptnst]] [[exp:=1]] <<DEATH >>[[record]] [[aes]]

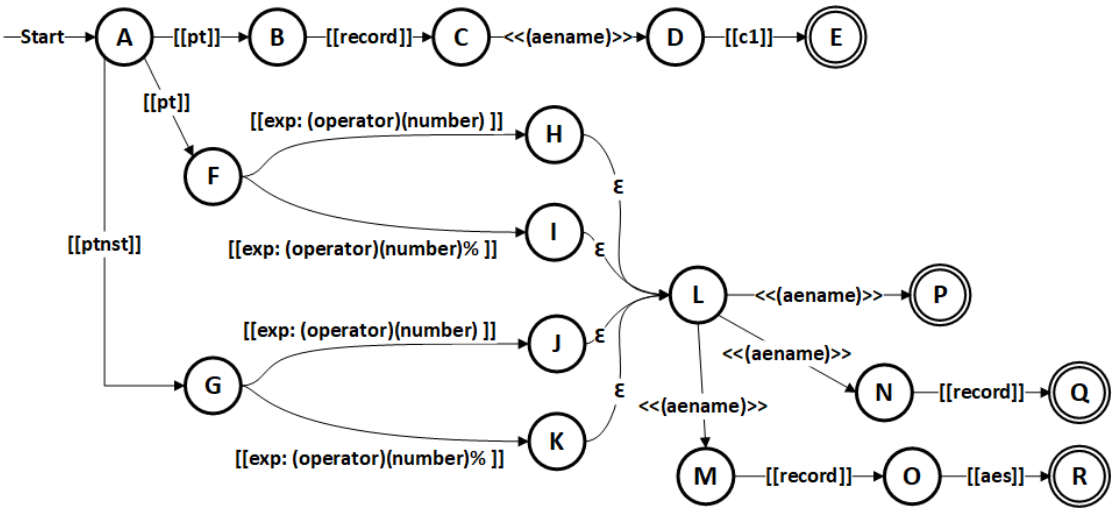


FIGURE 3.7: Automata Representation for Rule 4

Where possible values of terms enclosed in round brackets () in Figure 3.7 are given below: These terms represent specific variables or parameters that can take on a range of values.

- **Operator:** `>, <, and =`
- **Number:** Any decimal number
- **Ae Name:** Name of adverse event / disease

In order for this rule to match a statement, it needs to match keywords in one of the four expression groups, which are mentioned below.

Expressions Group 1

- `[[pt]]`
- `[[record]]`
- `<<Adverse Event Name>>`
- `[[c1]]`

Expressions Group 2

- `[[pt]]` or `[[ptnst]]`
- `[[expression]]`
- `<<Adverse Event Name>>`

Expressions Group 3

- `[[pt]]` or `[[ptnst]]`
- `[[expression]]`

- `<<Adverse Event Name>>`
- `[[record]]`

Expressions Group 4

- `[[pt]]` or `[[ptnst]]`
- `[[expression]]`
- `<<Adverse Event Name>>`
- `[[record]]`
- `[[aes]]`

A visual representation of this rule, in the form of an automaton, is presented in Figure 3.7. After rule matching process, the system proceeds to rule processing. This phase involves analysis whether the hypothesis logically follows from the CTR Premise. The CTR Premise text from CTR file relevant to this hypothesis is provided below.

CTR Premise

```
Primary Evidence:
{'Cohort1': {'Adverse Events': [{'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Death, not
                                 otherwise specified',
                                 'Patients': '0',
                                 'Percentage': '0.00',
                                 'Synonyms': '',
                                 'Total Patients': '9'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Pain, bone',
                                 'Patients': '1',
                                 'Percentage': '11.11',
                                 'Synonyms': '',
                                 'Total Patients': '9'}],
             'Percentage': 11.11,
             'Total': 1,
             'TotalPatients': 9},
```

```
 'Cohort2': {'Adverse Events': [{'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Death, not
                                 otherwise specified',
                                 'Patients': '1',
                                 'Percentage': '11.11',
                                 'Synonyms': '',
                                 'Total Patients': '9'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Pain, bone',
                                 'Patients': '0',
                                 'Percentage': '0.00',
                                 'Synonyms': '',
                                 'Total Patients': '9'}],
             'Percentage': 11.11,
             'Total': 1,
             'TotalPatients': 9}}
Secondary Evidence:
{'Cohort1': {'Adverse Events': [{'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Neutrophils/
                                 granulocytes (ANC/AGC)',
                                 'Patients': '1',
                                 'Percentage': '0.48',
                                 'Synonyms': '',
                                 'Total Patients': '208'}],
             'Percentage': 0.48,
             'Total': 1,
             'TotalPatients': 208},
 'Cohort2': {'Adverse Events': [{'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Neutrophils/
                                 granulocytes (ANC/AGC)',
                                 'Patients': '0',
                                 'Percentage': '0.00',
                                 'Total Patients': '201'}],
             'Percentage': 0.0,
             'Total': 0,
             'TotalPatients': 201}}
```


Hypothesis indicates that 1 death must exist in primary trial or secondary trial which is evident from CTR Premise that 1 death exist in Cohort 2 of Primary Trial. The results is entailment in both the proposed system and the gold standard.

Rule 5

Rule 5 matches keywords and adverse event name in hypothesis. Hypothesis and its pre-processed text are given below

> **Hypothesis:** The only types of Aes observed in patients from the secondary trial were Eyelid oedema, Upper gastrointestinal haemorrhage and Chest pain, no aes were recorded in the primary trial.
>
> **Processed Text:** [[aestypes]] [[record]] [[st]] <<EYELID OEDEMA >>, <<UPPER GASTROINTESTINAL HAEMORRHAGE >>and <<CHEST PAIN >>, no [[aes]] [[record]] [[pt]]

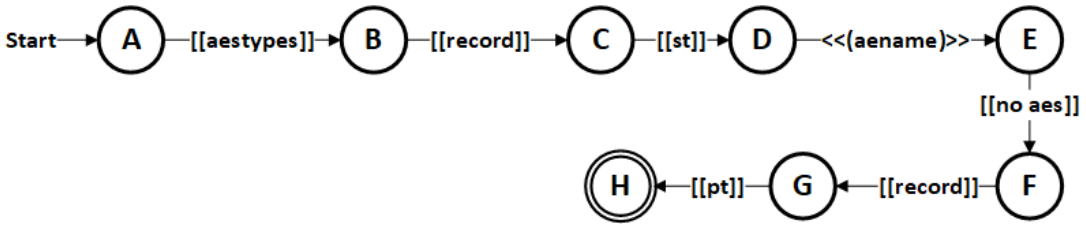


FIGURE 3.8: Automata Representation for Rule 5

where (aename) in Figure 3.8 will be the name of any adverse event / disease.

In order for this rule to match a statement, it needs to match following keywords.

- `[[aestypes]]`
- `[[record]]`
- `[[st]]`
- `<<Adverse Event Name>>`
- `[[no aes]]`
- `[[record]]`
- `[[pt]]`

Automata representation of this rule is shown in Figure 3.8. After rule matching, rule processor will find out whether the hypothesis entails the CTR Premise. The CTR Premise for this hypothesis is as under.

CTR Premise

```
Primary Evidence:
{'Cohort1': {'Adverse Events': [],
             'Percentage': 0.0,
             'Total': 0,
             'TotalPatients': 21},
 'Cohort2': {'Adverse Events': [],
             'Percentage': 0,
             'Total': 0,
             'TotalPatients': 0}}
Secondary Evidence:
{'Cohort1': {'Adverse Events': [{'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Anaemia',
                                 'Patients': '0',
                                 'Percentage': '0.00',
                                 'Synonyms': '',
                                 'Total Patients': '8'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Febrile neutropenia',
                                 'Patients': '0',
                                 'Percentage': '0.00',
                                 'Synonyms': '',
                                 'Total Patients': '8'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Polycythaemia',
                                 'Patients': '0',
                                 'Percentage': '0.00',
                                 'Synonyms': '',
                                 'Total Patients': '8'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Acute coronary syndrome',
                                 'Patients': '0',
                                 'Percentage': '0.00',
                                 'Synonyms': '',
                                 'Total Patients': '8'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Eyelid oedema',
                                 'Patients': '1',
                                 'Percentage': '12.50',
                                 'Synonyms': '',
```

```
                                'Total Patients': '8'},
                               {'Abbreviation': '',
                                'Hypernyms': '',
                                'Name': 'Constipation',
                                'Patients': '0',
                                'Percentage': '0.00',
                                'Synonyms': '',
                                'Total Patients': '8'},
                               {'Abbreviation': '',
                                'Hypernyms': '',
                                'Name': 'Diarrhoea',
                                'Patients': '0',
                                'Percentage': '0.00',
                                'Synonyms': '',
                                'Total Patients': '8'},
                               {'Abbreviation': '',
                                'Hypernyms': '',
                                'Name': 'Nausea',
                                'Patients': '0',
                                'Percentage': '0.00',
                                'Synonyms': '',
                                'Total Patients': '8'},
                               {'Abbreviation': '',
                                'Hypernyms': '',
                                'Name': 'Stomatitis',
                                'Patients': '0',
                                'Percentage': '0.00',
                                'Synonyms': '',
                                'Total Patients': '8'},
                               {'Abbreviation': '',
                                'Hypernyms': '',
                                'Name': 'Upper gastrointestinal
                                haemorrhage',
                                'Patients': '0',
                                'Percentage': '0.00',
                                'Synonyms': '',
                                'Total Patients': '8'},
                               {'Abbreviation': '',
                                'Hypernyms': '',
                                'Name': 'Vomiting',
                                'Patients': '0',
                                'Percentage': '0.00',
                                'Synonyms': '',
                                'Total Patients': '8'},
                               {'Abbreviation': '',
                                'Hypernyms': '',
                                'Name': 'Chest pain',
                                'Patients': '0',
                                'Percentage': '0.00',
                                'Synonyms': '',
                                'Total Patients': '8'}],
            'Percentage': 37.5,
```

```
            'Total': 3,
            'TotalPatients': 8},
 'Cohort2': {'Adverse Events': [{'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Anaemia',
                                 'Patients': '0',
                                 'Percentage': '0.00',
                                 'Synonyms': '',
                                 'Total Patients': '6'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Febrile neutropenia',
                                 'Patients': '0',
                                 'Percentage': '0.00',
                                 'Synonyms': '',
                                 'Total Patients': '6'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Polycythaemia',
                                 'Patients': '0',
                                 'Percentage': '0.00',
                                 'Synonyms': '',
                                 'Total Patients': '6'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Acute coronary syndrome',
                                 'Patients': '0',
                                 'Percentage': '0.00',
                                 'Synonyms': '',
                                 'Total Patients': '6'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Eyelid oedema',
                                 'Patients': '0',
                                 'Percentage': '0.00',
                                 'Synonyms': '',
                                 'Total Patients': '6'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Constipation',
                                 'Patients': '0',
                                 'Percentage': '0.00',
                                 'Synonyms': '',
                                 'Total Patients': '6'},
                                {'Abbreviation': '',
                                 'Hypernyms': '',
                                 'Name': 'Diarrhoea',
                                 'Patients': '0',
                                 'Percentage': '0.00',
                                 'Synonyms': '',
                                 'Total Patients': '6'},
                                {'Abbreviation': '',
```

```
                          'Hypernyms': '',
                          'Name': 'Nausea',
                          'Patients': '0',
                          'Percentage': '0.00',
                          'Synonyms': '',
                          'Total Patients': '6'},
                         {'Abbreviation': '',
                          'Hypernyms': '',
                          'Name': 'Stomatitis',
                          'Patients': '0',
                          'Percentage': '0.00',
                          'Synonyms': '',
                          'Total Patients': '6'},
                         {'Abbreviation': '',
                          'Hypernyms': '',
                          'Name': 'Upper gastrointestinal
                          haemorrhage',
                          'Patients': '0',
                          'Percentage': '0.00',
                          'Synonyms': '',
                          'Total Patients': '6'},
                         {'Abbreviation': '',
                          'Hypernyms': '',
                          'Name': 'Vomiting',
                          'Patients': '0',
                          'Percentage': '0.00',
                          'Synonyms': '',
                          'Total Patients': '6'},
                         {'Abbreviation': '',
                          'Hypernyms': '',
                          'Name': 'Chest pain',
                          'Patients': '1',
                          'Percentage': '16.67',
                          'Synonyms': '',
                          'Total Patients': '6'}],
'Percentage': 33.33,
'Total': 2,
'TotalPatients': 6}}
```


Hypothesis indicates that EYELID OEDEMA, UPPER GASTROINTESTINAL HAEMORRHAGE and CHEST PAIN must exist in secondary trial whereas no aes should exist in primary trial. It can be seen in CTR premise that second condition that no Aes should exist in primary trial is fulfilled. However, In cohort 1 of Secondary trial 1 patient exist for EYELID OEDEMA where as no patient exist for UPPER GASTROINTESTINAL HAEMORRHAGE and CHEST PAIN, whereas in cohort 2 of Secondary trial 1 patient exist for CHEST PAIN where

as no patient exist for UPPER GASTROINTESTINAL HAEMORRHAGE and EYELID OEDEMA. Which makes it a contradiction which is same as mentioned in the gold standard.

Algorithm 1 Rule Matching

```
1: Input: S  // Input statement
2: Output: Rule number if a match is found, otherwise no rule matched
3:
4: Rules[ ][ ]  // regex for all rules
5:
6: eS ← preprocessing(S)
7: for each R in Rules[ ][ ] do
8:    results ← apply_regex(eS, R)
9:    totalKeywords ← len(results)
10:   foundKeywords ← findKeywordOccurances(eS)
11:   if totalKeywords = foundKeywords then
12:      return  R['RuleNo']
13:   end if
14: end for
15: Return: No rule matched
```

The rule matching algorithm begins by accepting an input statement, denoted as S, and aims to determine if it matches any predefined rules. If a match is found, the algorithm returns the corresponding rule number; otherwise, it indicates that no rule was matched. The rules are stored in a 2D array called 'Rules', where each row represents a different rule with associated regex patterns.

The algorithm first preprocesses the input statement S to clean and standardize it, resulting in eS. It then iterates through each rule in the 'Rules' array, applying the corresponding regex to the preprocessed input statement. The algorithm counts the total number of keywords matched by the regex and also identifies how many distinct keywords defined in the rule appear in eS.

By comparing these two counts, the algorithm checks if all required keywords were found in the input statement. If they match, it returns the rule number associated with the matching rule. If no matches are found after iterating through all rules, the algorithm concludes by stating that no rules were matched.

Algorithm 2 Rule Processing

```
Input:
  eS: string
  ruleNo: integer
Output:
  result = "Entailment" or "Contradiction"
expNum ← getNumberInExpression(eS)
if ruleNo = 1 then
  if re.search("[[exp:<+%]]", eS) then
    if primaryEvidence['Cohort1']['Percentage'] < expNum then
      return "Entailment"
    end if
  end if
else if ruleNo = 2 then
  if SecAesMoreThanPriAes () then
    return "Entailment"
  end if
else if ruleNo = 3 then
  if primaryEvidence['Cohort1']['Total'] = primaryEvidence['Cohort2']['Total'] then
    return "Entailment"
  end if
else if ruleNo = 4 then
  i ← 0
  matchAes ← findAesInStmt(eS)
  for each match in matchAes do
    if PtC1NotContainsAe(match) and PtC2ContainAes(match) then
      return "Entailment"
    end if
  end for
else if ruleNo = 5 then
  matchAes ← findAesInStmt(eS)
  if (primaryEvidence['Cohort1']['Total'] = 0) and (primaryEvidence['Cohort2']['Total'] = 0) then
    flagFoundInPt ← True
  end if
  numFound ← 0
  for each match in matchAes do
    if PtC1ContainsAe(match) then
      numFound ← numFound + 1
    end if
  end for
  if flagFoundInPt and totalAes = numFound then
    return "Entailment"
  end if
end if
return "Contradiction"
```

3.2.4 Dataset

The NLI4CT dataset is prepared by experts in the clinical domain, including clinical trial organizers and research oncologists from the Cancer Research UK Manchester Institute, as well as members of the Digital Experimental Cancer Medicine Team. [53]

The dataset consisted of a collection of breast cancer clinical trial reports and designed to enhance research in natural language inference with a focus on clinical trial reports. It features a comprehensive collection of 2,400 annotated statements, each paired with corresponding labels, CTRs, and evidence. The dataset is organized into distinct subsets: 1,700 samples for training, 500 for testing, and 200 for development. Each CTR is broken down into four key sections—eligibility criteria, intervention, results, and adverse events—where the statements make claims regarding the information contained in these sections. The primary task is to assess the inference relation, such as entailment or contradiction, between the CTR and the provided statement.

Each CTR can be organized into four distinct sections

- **Eligibility Criteria:** Defines the conditions that patients must meet to qualify for participation in the clinical trial.
- **Intervention:** Details the type, dosage, frequency, and duration of the treatments being investigated.
- **Results:** Provides information on the number of participants, outcome measures, units of measurement, and the overall results.
- **Adverse Events:** Documents the signs and symptoms observed in patients throughout the clinical trial.

For the competition the test and development sets were modified with a variety of interventions, targeting numerical reasoning, vocabulary, syntax, and semantics,

to systematically investigate the consistency, robustness, and faithfulness of NLI models in clinical settings.

To address specific challenges faced by NLI models, the dataset includes enriched interventions in the test and development sets, targeting aspects such as numerical reasoning, vocabulary and syntax, and semantic complexity. The goal is to bring advancements in developing robust and accurate NLI models for extracting and interpreting medical evidence from a large corpus of CTRs. To support further research, the dataset, competition leaderboard, website, and code for replicating baseline experiments are all publicly available, making it a valuable resource for those working on medical evidence extraction from clinical trial reports.

3.2.5 Correction of Annoatation Error

Our rule-based system encountered a false case, prompting a thorough investigation into the underlying causes. This in-depth analysis revealed that the issue stemmed from an annotation error within the dataset. Such errors can significantly impact the performance and accuracy of any machine learning model, as they may lead to misinterpretations of the data.

Our rule-based system rigorously analyzed the dataset and accurately identified instances of erroneous annotations. By rectifying these errors, we have significantly improved the system's reliability and overall performance. By ensuring the accuracy and consistency of the dataset, we have minimized the likelihood of false positive cases and strengthened the validity of the results produced by the rule-based system. Subsequent to this dataset correction, we re-evaluated the results of Zhou et al.'s system, which subsequently decreased to 0.854.

Chapter 4

Experimentation

The experiments for this study were conducted using free version of Google Colab, a cloud-based platform that provides a convenient and accessible environment for running Jupyter notebooks. Google colab was using Python 3.10.12 at the time of experimentation. Google Colab typically provides around 12 GB of RAM and access to a standard CPU, such as an Intel Xeon or a similar model. For storage, users have approximately 50 GB of local temporary disk space.

The packages necessary for implementing this code are as follows

- `import json`
- `import nltk`
- `from nltk.corpus import stopwords`
- `from nltk.tokenize import word_tokenize`
- `from word2number import w2n`
- `from google.colab import drive`
- `import pandas as pd`
- `import os`
- `import re`
- `import copy`
- `import pprint`

4.1 Evaluation

4.1.1 Evaluation Measure

The performance of NLI systems is assessed using F1 Score, also known as the F measure. It is the harmonic mean of precision and recall. F1 score is a valuable metric for assessing the performance of classification models by considering both precision and recall.

$$\text{F1 Score} = \frac{2TP}{2TP + FP + FN}$$

where:

- TP = True Positives
- FP = False Positives
- FN = False Negatives

4.1.2 Results

Table 4.2 showcases the significant performance improvement achieved by our proposed system, which combines a deep learning model with a rule-based system. This hybrid approach yielded an impressive F1 score of 0.870 on the NLI4CT dataset, surpassing state-of-the-art systems like Zhou's [54] and Kamal's [81]. These systems, while strong performers, achieved F1 scores of 0.856 and 0.834, respectively, on all sections of CTRs.

By leveraging the strengths of both deep learning and rule-based techniques, our proposed system effectively addresses the limitations of traditional approaches. The deep learning component enables the model to capture complex patterns and dependencies within the data, while the rule-based component provides a structured and interpretable way to incorporate domain-specific knowledge. Zhou's system encountered failures in 8 cases, which was found by comparing with Gold Standard.

TABLE 4.1: Predictions Comparison between Zhou's System and Proposed System

Sr #	Statement	Zhou	Our System
1	AEs were not recorded for the primary trial or the secondary trial.	T	T
2	Less than 5% of patients in the primary trial suffered AEs	F	T
4	Less than 10 patients in the primary trial suffered AEs	F	T
5	Over 97% of patients in the secondary trial and the primary trial did not suffer any adverse events	T	F
6	Cohort 1 of the primary trial did not report any AEs	T	T
7	In both the primary trial and the secondary trial there were several adverse events which occurred in more than 30% of participants.	T	T
8	The secondary trial recorded 11 more AEs than the primary trial	T	T
9	The secondary trial recorded slightly more total adverse events in its patient cohorts than the primary trial.	F	T
10	The same number of AEs were reported for both cohorts in the primary trial	T	T
11	The primary trial and the secondary trial recorded the same total number of adverse events in their patient cohorts.	F	T
12	The primary trial and the secondary trial recorded the same number of adverse events in their cohorts.	T	T
13	All Infections and Fever cases in the primary trial were for patients in cohort 1.	F	T
14	All Infections and Infestations cases in the primary trial were for patients in cohort 1.	F	T
15	Across both the primary trial and the secondary trial only one death was recorded in the adverse events.	T	T
16	Over 15% of patients in the primary trial and the secondary trial suffered from infections during the study period	T	T
17	There was at least 1 case of infection in both the primary trial and the secondary trial	T	T
18	Across both the primary trial and the secondary trial over 10 deaths were recorded in the adverse events.	T	T
19	36.36% of the primary trial patients suffered an increase in Blood bilirubin.	T	T
20	25% of patients in the primary trial suffer Increased pleural effusion and Rapid disease progression	T	T
21	25% of cohort 2 patients in the primary trial suffer Increased pleural effusion	T	T
22	There was at least 1 recorded gastro-intestinal adverse event in the primary trial	T	T
23	The only types of AEs observed in patients from the secondary trial were Eyelid oedema, Upper gastrointestinal haemorrhage and Chest pain, no AEs were recorded in the primary trial	F	T

The proposed hybrid system was evaluated on a test dataset of 23 cases that matched the identified failure patterns. The system successfully predicted 22 out of these 23 cases, significantly outperforming the previous state-of-the-art system, which correctly predicted only 15 cases as shown in Table 4.1. This improvement highlights the effectiveness of our approach in addressing the limitations of existing methods and demonstrates the potential of combining deep learning with rule-based techniques for natural language inference tasks.

TABLE 4.2: Top Performer on Textual Entailment for NLI4CT Dataset

Sr. No.	**Reference**	**F1 Score**
1	Proposed System	87.00%
2	Zhou et al., 2023 [54]	85.40%
3	Kanakarajan et al., 2023 [81]	83.40%
4	Liu et al., 2024 [77]	80.00%
5	Guimarães et al., 2024 [83]	80.00%
6	Vladika et al., 2023 [68]	79.80%
7	Lee et al., 2024 [84]	77.90%
8	Wang et al., 2023 [71]	76.40%
9	Chakraborty, 2024 [78]	76.00%
10	Smilga, 2024 [82]	76.00%
11	Chen et al., 2019 [87]	70.90%
12	Alameldin et al., 2023 [73]	70.50%
13	Rajamanickam et al., 2023 [69]	70.10%
14	Bevan et al., 2023 [74]	69.50%
15	Zhao et al., 2023 [79]	67.90%
16	Pahwa et al., 2023 [80]	67.90%
17	Feng et al., 2023 [75]	67.90%
18	Alissa et al., 2023 [76]	67.00%
19	Noor Mohamed et al., 2023 [85]	66.70%

Chapter 5

Conclusion and Future Work

In this thesis, a thorough review of the existing literature on NLP/NLI systems has been conducted, accompanied by appropriate critical comments. The review indicates that most published research predominantly employs deep learning-based approaches especially Transformer based systems. Deep learning has revolutionized many fields by providing powerful tools for solving complex problems, particularly those involving large-scale data and intricate patterns. However, still there are cases which are overlooked by Deep Learning system, this problems can be effectively addressed with simpler algorithms. This raises the question: Is it always necessary to use deep learning, or can traditional methods be sufficient?

The answer lies in Hybrid system that utilizes deep learning approach with rule-based heuristics. This approach combines the strength of deep learning with a rule-based system to enhance overall performance. Rule-based approach is usually capable of solving specific problems efficiently, requiring less time and fewer resources. While rule-based systems may lack the flexibility and scalability of deep learning models, they often deliver superior results in NLP tasks where the structure of language is relatively fixed and well-defined. It was observed that the input statements where the deep learning system failed, they contain common words and follow fixed sentence structure. These recurring patterns were identified, and corresponding rules were developed to detect the patterns. This method proved effective, raising the F1 score to 0.870 marking a significant improvement over the previous system.

Additionally, the rule-based system is deterministic, providing faster results with greater efficiency. We employed Leading Deep Learning system [54] which shown efficacy in most of the cases, while for failure cases we employed Rule-based system which excel in handling specific types of inference scenarios, particularly where specific linguistic pattern can be seen. However, there is potential to expand the rule set to cover a broader range of linguistic phenomena and to improve External Knowledge Base.Our Rule-based system also identified an annotation error in dataset, which we corrected and updated F1 score of DL system.

Future work could focus on incorporating additional rules and knowledge base to handle other sections of CTRs i.e. Intervention, Results and Eligibility Criteria. Furthermore, The proposed system can be tailored for specific types of textual entailment scenarios. Adapting the system to work across different domains (e.g., legal texts, technical documentation) could be an interesting avenue for future research.

To summarize, Deep Learning system along with Rule-based systems have demonstrated significant strengths in certain areas, particularly where explicit patterns can be captured in rules. They offer high interpretability and reliability in scenarios where precision is critical. However, the future lies in these kind of hybrid systems that combine the best of both techniques, with the transparency of rule-based methods and the flexibility of machine learning models to create more robust and adaptable NLI systems.

Bibliography

[1] I. M. S. Putra, D. Siahaan, and A. Saikhu, "Recognizing textual entailment: A review of resources, approaches, applications, and challenges," *ICT Express*, vol. 10, no. 1, pp. 132–155, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405959523001145

[2] M. Jullien, M. Valentino, H. Frost, P. O'regan, D. Landers, and A. Freitas, "SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 2216–2226. [Online]. Available: https://aclanthology.org/2023.semeval-1.307

[3] M. Fazelnia, V. Koscinski, S. Herzog, and M. Mirakhorli, " Lessons from the Use of Natural Language Inference (NLI) in Requirements Engineering Tasks ," in *2024 IEEE 32nd International Requirements Engineering Conference (RE)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2024, pp. 103–115. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/RE59067.2024.00020

[4] R. Gubelmann, I. Katis, C. Niklaus, and S. Handschuh, "Capturing the varieties of natural language inference: A systematic survey of existing datasets and two novel benchmarks," *Journal of Logic, Language and Information*, vol. 33, no. 1, pp. 21–48, 2024.

[5] M. Valentino, "Explanation-based scientific natural language inference," Ph.D. dissertation, The University of Manchester, 2022.

[6] U.S. National Library of Medicine. (2024) Clinicaltrials.gov. Accessed: 2024-09-02. [Online]. Available: https://clinicaltrials.gov/

[7] S. Kumar, P. Kaur, and A. Gosain, "A comprehensive survey on ensemble methods," in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, IEEE. IEEE, 2022, pp. 1–7.

[8] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, 2023.

[9] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99 129–99 149, 2022.

[10] J. Jia, W. Liang, and Y. Liang, "A review of hybrid and ensemble in deep learning for natural language processing," *arXiv preprint arXiv:2312.05589*, 2023.

[11] Y. Kim and S. M. Meystre, "Ensemble method–based extraction of medication and related information from clinical texts," *Journal of the American Medical Informatics Association*, vol. 27, no. 1, pp. 31–38, 2020.

[12] B. Alekhya and R. Sasikumar, "An ensemble approach for healthcare application and diagnosis using natural language processing," *Cognitive Neurodynamics*, vol. 16, no. 5, pp. 1203–1220, 2022.

[13] R. Baradaran and H. Amirkhani, "Ensemble learning-based approach for improving generalization capability of machine reading comprehension systems," *Neurocomputing*, vol. 466, pp. 229–242, 2021.

[14] Z. Al-Makhadmeh and A. Tolba, "Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach," *Computing*, vol. 102, no. 2, pp. 501–522, 2020.

[15] H. Zhao, J. Zhu, and W. Deng, "A new weighted ensemble model-based method for text implication recognition," *Multimedia Tools and Applications*, pp. 1–16, 2024.

[16] S. Jaradat, R. Nayak, A. Paz, and M. Elhenawy, "Ensemble learning with pre-trained transformers for crash severity classification: A deep nlp approach," *Algorithms*, vol. 17, no. 7, p. 284, 2024.

[17] J. Fattahi and M. Mejri, "Spaml: a bimodal ensemble learning spam detector based on nlp techniques," in *2021 IEEE 5th international conference on cryptography, security and privacy (CSP)*. IEEE, 2021, pp. 107–112.

[18] L. Ansari, S. Ji, Q. Chen, and E. Cambria, "Ensemble hybrid learning methods for automated depression detection," *IEEE transactions on computational social systems*, vol. 10, no. 1, pp. 211–219, 2022.

[19] A. Mathew, P. Amudha, and S. Sivakumari, "Deep learning techniques: an overview," *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020*, pp. 599–608, 2021.

[20] S. Lawrence, C. L. Giles, and S. Fong, "Natural language grammatical inference with recurrent neural networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 12, no. 1, pp. 126–140, 2000.

[21] N. Jiang and M.-C. de Marneffe, "Evaluating BERT for natural language inference: A case study on the CommitmentBank," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6086–6091. [Online]. Available: https://aclanthology.org/D19-1630

[22] K. S. Kalyan, "A survey of gpt-3 family large language models including chatgpt and gpt-4," *Natural Language Processing Journal*, p. 100048, 2023.

[23] L. Yao and Y. Guan, "An improved lstm structure for natural language processing," in *2018 IEEE international conference of safety produce informatization (IICSPI)*, IEEE. IEEE, 2018, pp. 565–569.

[24] O. Lokshyn, "Natural language inference: An overview," 2024, towards Data Science. [Online]. Available: https://towardsdatascience.com/natural-language-inference-an-overview-5e2d6c5b4f32

[25] Z. Chen, Q. Gao, and L. S. Moss, "NeuralLog: Natural language inference with joint neural and logical reasoning," in *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*. Online: Association for Computational Linguistics, Aug. 2021, pp. 78–88. [Online]. Available: https://aclanthology.org/2021.starsem-1.7

[26] Y. Kim, M. Jang, and J. Allan, "Explaining text matching on neural natural language inference," *ACM Transactions on Information Systems*, vol. 38, no. 4, pp. 1–23, 2020.

[27] L. Von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy *et al.*, "Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 614–633, 2021.

[28] L. Patel, T. Shukla, X. Huang, D. W. Ussery, and S. Wang, "Machine learning methods in drug discovery," *Molecules*, vol. 25, no. 22, p. 5277, 2020.

[29] B. Magnini, A. Lavelli, and S. Magnolini, "Comparing machine learning and deep learning approaches on NLP tasks for the Italian language," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 2110–2119. [Online]. Available: https://aclanthology.org/2020.lrec-1.259

[30] G. Engin, B. Aksoyer, M. Avdagic, D. Bozanlı, U. Hanay, D. Maden, and G. Ertek, "Rule-based expert systems for supporting university students," *Procedia Computer Science*, vol. 31, pp. 22–31, 2014.

[31] I. Mikulić, D. Lisjak, and N. Štefanić, "A rule-based system for human performance evaluation: A case study," *Applied Sciences*, vol. 11, no. 7, p. 2904, 2021.

[32] R. Patil, S. Boit, V. Gudivada, and J. Nandigam, "A survey of text representation and embedding techniques in nlp," *IEEE Access*, vol. 11, pp. 36 120–36 146, 2023.

[33] R. Gabud, P. Lapitan, V. Mariano, E. Mendoza, N. Pampolina, M. A. A. Clariño, and R. Batista-Navarro, "A hybrid of rule-based and transformer-based approaches for relation extraction in biodiversity literature," in *Proceedings of the 2nd Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*, M. Surdeanu, E. Riloff, L. Chiticariu, D. Frietag, G. Hahn-Powell, C. T. Morrison, E. Noriega-Atala, R. Sharp, and M. Valenzuela-Escarcega, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 103–113. [Online]. Available: https://aclanthology.org/2023.pandl-1.10

[34] B. Silva, F. Hak, T. Guimarães, M. Manuel, and M. F. Santos, "Rule-based system for effective clinical decision support," *Procedia Computer Science*, vol. 220, pp. 880–885, 2023.

[35] A. Fantechi, S. Gnesi, and L. Semini, "Rule-based NLP vs chatgpt in ambiguity detection, a preliminary study," in *Joint Proceedings of REFSQ-2023 Workshops, Doctoral Symposium, Posters & Tools Track and Journal Early Feedback co-located with the 28th International Conference on Requirements Engineering: Foundation for Software Quality (REFSQ 2023), Barcelona, Catalunya, Spain, April 17-20, 2023*, ser. CEUR Workshop Proceedings, vol. 3378. CEUR-WS.org, 2023. [Online]. Available: https://ceur-ws.org/Vol-3378/NLP4RE-paper1.pdf

[36] S. S. Vedaei, A. Fotovvat, M. R. Mohebbian, G. M. E. Rahman, K. A. Wahid, P. Babyn, H. R. Marateb, M. Mansourian, and R. Sami, "Covid-safe: An iot-based system for automated health monitoring and surveillance in post-pandemic life," *IEEE Access*, vol. 8, pp. 188 538–188 551, 2020.

[37] GeeksforGeeks, "Rule-based approach in nlp," 2024, accessed: 2024-07-16. [Online]. Available: https://geeksforgeeks.org/rule-based-approach-in-nlp/

[38] R. Mumtaz and M. Qadir, "Custner: A rule-based named-entity recognizer with improved recall," *International Journal on Semantic Web and Information Systems*, vol. 16, no. 3, pp. 110–127, 2020.

[39] A.-L. Kalouli, R. Crouch, and V. de Paiva, "Hy-NLI: a hybrid system for natural language inference," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 5235–5249. [Online]. Available: https://aclanthology.org/2020.coling-main.459

[40] A. Willis, H. Yang, and A. De Roeck, "A generalised hybrid architecture for NLP," in *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, N. Grabar, M. Dupuch, A. Périnet, and T. Hamon, Eds. Avignon, France: Association for Computational Linguistics, Apr. 2012, pp. 97–105. [Online]. Available: https://aclanthology.org/W12-0513

[41] M. Jullien, M. Valentino, H. Frost, P. O'Regan, D. Landers, and A. Freitas, "NLI4CT: Multi-evidence natural language inference for clinical trial reports," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 16 745–16 764. [Online]. Available: https://aclanthology.org/2023.emnlp-main.1041

[42] P. Lu, L. Qiu, W. Yu, S. Welleck, and K.-W. Chang, "A survey of deep learning for mathematical reasoning," in *Proceedings of the 61st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 14 605–14 631. [Online]. Available: https://aclanthology.org/2023.acl-long.817

[43] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," pp. 632–642, Sep. 2015. [Online]. Available: https://aclanthology.org/D15-1075

[44] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1112–1122. [Online]. Available: https://aclanthology.org/N18-1101

[45] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee *et al.*, "Natural questions: a benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, 2019.

[46] L. Sharma, L. Graesser, N. Nangia, and U. Evci, "Natural language understanding with the quora question pairs dataset," *arXiv preprint arXiv:1907.01041*, 2019.

[47] A. Wang, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.

[48] N. Xie, F. Lai, D. Doran, and A. Kadav, "Visual entailment: A novel task for fine-grained image understanding," *arXiv preprint arXiv:1901.06706*, 2019.

[49] L. Bentivogli, R. Bernardi, M. Marelli, S. Menini, M. Baroni, and R. Zamparelli, "Sick through the semeval glasses. lesson learned from the evaluation of compositional distributional semantic models on full sentences through

semantic relatedness and textual entailment," *Language Resources and Evaluation*, vol. 50, pp. 95–124, 2016.

[50] K. Balog, P. Serdyukov, and A. P. de Vries, "Overview of the trec 2011 entity track." in *TREC*, vol. 2011, 2011, p. 11.

[51] W. Yin, D. Radev, and C. Xiong, "DocNLI: A large-scale dataset for document-level natural language inference," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 4913–4922. [Online]. Available: https://aclanthology.org/2021.findings-acl.435

[52] Y. Koreeda and C. Manning, "ContractNLI: A dataset for document-level natural language inference for contracts," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1907–1919. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.164

[53] S. Organization, "Semeval — international workshop on semantic evaluation," 2023, website for the International Workshop on Semantic Evaluation. [Online]. Available: http://www.semeval.org/

[54] Y. Zhou, Z. Jin, M. Li, M. Li, X. Liu, X. You, and J. Wu, "THiFLY research at SemEval-2023 task 7: A multi-granularity system for CTR-based textual entailment and evidence retrieval," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1681–1690. [Online]. Available: https://aclanthology.org/2023.semeval-1.234

[55] J. Romero, S. Chiang, and D. M. Goldenholz, "Can machine learning improve randomized clinical trial analysis?" *Seizure*, vol. 91, pp. 499–502, 2021.

[56] R. Mahendra, D. Spina, L. Cavedon, and K. Verspoor, "Do numbers matter? types and prevalence of numbers in clinical texts," in *Proceedings of the 23rd*

Workshop on Biomedical Natural Language Processing, D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, and J. Tsujii, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 409–415. [Online]. Available: https://aclanthology.org/2024.bionlp-1.32

[57] n2c2, "n2c2 nlp research data sets," 2024, accessed: 2024-08-02. [Online]. Available: https://harvard.edu

[58] EBM-NLP Dataset, "Ebm-nlp dataset," 2024, accessed: 2024-08-02. [Online]. Available: https://paperswithcode.com/dataset/ebm-nlp

[59] Y. Gao, D. Dligach, L. Christensen, S. Tesch, R. Laffin, D. Xu, T. Miller, O. Uzuner, M. Churpek, and M. Afshar, "A scoping review of publicly available language tasks in clinical natural language processing," *Journal of the American Medical Informatics Association*, vol. 29, no. 10, pp. 1797–1806, 2022.

[60] Text REtrieval Conference (TREC). (2021) Trec 2021 clinical decision support track. Accessed: 2024-08-18. [Online]. Available: https://www.trec-cds.org/2021.html

[61] L. Biester, V. Joopudi, and B. Dandala, "Ibm@ trec clinical trials track 2021," in *Proceedings of TREC*, 2021.

[62] J. DeYoung, E. Lehman, B. Nye, I. Marshall, and B. C. Wallace, "Evidence inference 2.0: More data, better models," pp. 123–132, Jul. 2020. [Online]. Available: https://aclanthology.org/2020.bionlp-1.13

[63] A. Romanov and C. Shivade, "Lessons from natural language inference in the clinical domain," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 1586–1596. [Online]. Available: https://aclanthology.org/D18-1187

[64] S. Kierner, J. Kucharski, and Z. Kierner, "Taxonomy of hybrid architectures involving rule-based reasoning and machine learning in clinical decision systems: A scoping review," *Journal of Biomedical Informatics*, vol. 144, p. 104428, 2023.

[65] P. Ray and A. Chakrabarti, "A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis," *Applied Computing and Informatics*, vol. 18, no. 1/2, pp. 163–178, 2022.

[66] L. Chen, Y. Gu, X. Ji, C. Lou, Z. Sun, H. Li, Y. Gao, and Y. Huang, "Clinical trial cohort selection based on multi-level rule-based natural language processing system," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1218–1226, 2019.

[67] M. Wang, A. Agrawal, N. Rogers, V. John, and T. Thyvalikakath, "Rule-based text classification of dental diagnosis," in *MEDINFO 2023—The Future Is Accessible*. IOS Press, 2024, pp. 624–628.

[68] J. Vladika and F. Matthes, "Sebis at SemEval-2023 task 7: A joint system for natural language inference and evidence retrieval from clinical trial reports," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1863–1870. [Online]. Available: https://aclanthology.org/2023.semeval-1.257

[69] S. Rajamanickam and K. Rajaraman, "I2R at SemEval-2023 task 7: Explanations-driven ensemble approach for natural language inference over clinical trial data," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1630–1635. [Online]. Available: https://aclanthology.org/2023.semeval-1.226

[70] C. Takehana, D. Lim, E. Kurtulus, R. Iyer, E. Tanimura, P. Aggarwal, M. Cantillon, A. Yu, S. Khan, and N. Chi, "Stanford MLab at SemEval

2023 task 7: Neural methods for clinical trial report NLI," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1769–1775. [Online]. Available: https://aclanthology.org/2023.semeval-1.245

[71] W. Wang, B. Xu, T. Fang, L. Zhang, and Y. Song, "KnowComp at SemEval-2023 task 7: Fine-tuning pre-trained language models for clinical trial entailment identification," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1–9. [Online]. Available: https://aclanthology.org/2023.semeval-1.1

[72] M. Volosincu, C. Lupu, D. Trandabat, and D. Gifu, "FII SMART at SemEval 2023 task7: Multi-evidence natural language inference for clinical trial data," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 212–220. [Online]. Available: https://aclanthology.org/2023.semeval-1.30

[73] A. Alameldin and A. Williamson, "Clemson NLP at SemEval-2023 task 7: Applying GatorTron to multi-evidence clinical NLI," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1598–1602. [Online]. Available: https://aclanthology.org/2023.semeval-1.220

[74] R. Bevan, O. Turbitt, and M. Aboshokor, "MDC at SemEval-2023 task 7: Fine-tuning transformers for textual entailment prediction and evidence retrieval in clinical trials," in *Proceedings of the 17th International Workshop*

on Semantic Evaluation (SemEval-2023), A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1287–1292. [Online]. Available: https://aclanthology.org/2023.semeval-1.179

[75] C. Feng, J. Wang, and X. Zhang, "YNU-HPCC at SemEval-2023 task7: Multi-evidence natural language inference for clinical trial data based a BioBERT model," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 664–670. [Online]. Available: https://aclanthology.org/2023.semeval-1.91

[76] K. Alissa and M. Abdullah, "JUST-KM at SemEval-2023 task 7: Multi-evidence natural language inference using role-based double roberta-large," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 447–452. [Online]. Available: https://aclanthology.org/2023.semeval-1.61

[77] J. Liu and S. Thoma, "FZI-WIM at SemEval-2024 task 2: Self-consistent CoT for complex NLI in biomedical domain," pp. 1269–1279, Jun. 2024. [Online]. Available: https://aclanthology.org/2024.semeval-1.184

[78] A. Chakraborty, "RGAT at SemEval-2024 task 2: Biomedical natural language inference using graph attention network," in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 116–122. [Online]. Available: https://aclanthology.org/2024.semeval-1.19

[79] X. Zhao, M. Zhang, M. Ma, C. Su, Y. Liu, M. Wang, X. Qiao, J. Guo, Y. Li, and W. Ma, "HW-TSC at SemEval-2023 task 7: Exploring the natural language inference capabilities of ChatGPT and pre-trained language model for clinical trial," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1603–1608. [Online]. Available: https://aclanthology.org/2023.semeval-1.221

[80] B. Pahwa and B. Pahwa, "BpHigh at SemEval-2023 task 7: Can fine-tuned cross-encoders outperform GPT-3.5 in NLI tasks on clinical trial data?" in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1936–1944. [Online]. Available: https://aclanthology.org/2023.semeval-1.266

[81] K. R. Kanakarajan and M. Sankarasubbu, "Saama AI research at SemEval-2023 task 7: Exploring the capabilities of flan-t5 for multi-evidence natural language inference in clinical trial data," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 995–1003. [Online]. Available: https://aclanthology.org/2023.semeval-1.137

[82] V. Smilga and H. Alabiad, "TüDuo at SemEval-2024 task 2: Flan-t5 and data augmentation for biomedical NLI," in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 737–744. [Online]. Available: https://aclanthology.org/2024.semeval-1.106

[83] A. Guimarães, B. Martins, and J. Magalhães, "Lisbon computational linguists at SemEval-2024 task 2: Using a mistral-7B model and data augmentation," pp. 1280–1287, Jun. 2024. [Online]. Available: https://aclanthology.org/2024.semeval-1.185

[84] L.-h. Lee, C.-y. Chiou, and T.-m. Lin, "NYCU-NLP at SemEval-2024 task 2: Aggregating large language models in biomedical natural language inference for clinical trials," in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 1455–1462. [Online]. Available: https://aclanthology.org/2024.semeval-1.209

[85] S. S. Noor Mohamed and K. Srinivasan, "SSNSheerinKavitha at SemEval-2023 task 7: Semantic rule based label prediction using TF-IDF and BM25 techniques," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 950–957. [Online]. Available: https://aclanthology.org/2023.semeval-1.131

[86] M. Neves, "Bf3R at SemEval-2023 task 7: a text similarity model for textual entailment and evidence retrieval in clinical trials and animal studies," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 125–129. [Online]. Available: https://aclanthology.org/2023.semeval-1.17

[87] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li, X. Zhou, and W. Wang, "Tabfact: A large-scale dataset for table-based fact verification," *arXiv preprint*, 2019.

[88] A. Corrêa Dias, F. Dias, H. Moreira, V. Moreira, and J. L. Comba, "Team INF-UFRGS at SemEval-2023 task 7: Supervised contrastive learning for pair-level sentence classification and evidence retrieval," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 700–706. [Online]. Available: https://aclanthology.org/2023.semeval-1.96

[89] S. I. R. Conceição, D. F. Sousa, P. Silvestre, and F. M. Couto, "lasigeBioTM at SemEval-2023 task 7: Improving natural language inference baseline systems with domain ontologies," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10–15. [Online]. Available: https://aclanthology.org/2023.semeval-1.2

[90] R. Katiyar. (2023) What is rule engine? Accessed: 2024-08-18. [Online]. Available: https://medium.com/@er.rameshkatiyar/what-is-rule-engine-86ea759ad97d

Appendix A

Some examples of Clinical Trial Reports from NLI4CT Dataset

Example 1

Clinical Trial ID

- Clinical Trial ID: `NCT00005908`

Interventions

- **INTERVENTION 1:** Dose A-Cohort 1-Arm 1-Docetaxel & Capecitabine
  - Docetaxel 75 $\mathrm{mg/m^2}$ intravenous on Day 1, capecitabine 1000 $\mathrm{mg/m^2}$ orally twice daily from Day 2 to Day 15 for 4 cycles. Once the dose was deemed to be too toxic, subsequent patients were enrolled on Dose B.
- **INTERVENTION 2:** Dose B-Cohort 2-Arm 2 Reduced Dose-Docetaxel & Capecitabine
  - Docetaxel 60 $\mathrm{mg/m^2}$ intravenous on Day 1, capecitabine 937.5 $\mathrm{mg/m^2}$ orally twice daily from Day 2 to Day 15.

Eligibility

- **INCLUSION CRITERIA:**

  - Stage II or III breast cancer with a tumor size of greater than 2 cm. Patients with a previous biopsy are eligible provided adequate tumor tissue remains for biopsy in this study.
  - At least 18 years of age.
  - Adequate hematopoietic function as defined by absolute neutrophil count greater than 1200 mm^{-3} and platelet count greater than 100,000 mm^{-3}.
  - Adequate renal function as defined by creatinine less than 1.6 mg/dL.
  - Adequate hepatic function as defined by total bilirubin less than 1.4 mg/dL and serum glutamic oxaloacetic transaminase (SGOT)/serum glutamic pyruvic transaminase (SGPT) less than 1.5 times the upper limit of normal and alkaline phosphatase less than 2.5 times the upper limit of normal.
  - Zubrod Performance status 0-2.

- **EXCLUSION CRITERIA:**
  - Medical or psychiatric condition that, in the opinion of the Principal Investigator, would preclude chemotherapy administration. Patients may be evaluated by psychiatry or medical subspecialties as appropriate.
  - Pregnant or lactating women.
  - Known bleeding disorders.
  - Hypersensitivity to Tween 80 (Polysorbate).
  - Cardiac ejection fraction below normal limits, myocardial infarction within the past 12 months, or symptomatic arrhythmia requiring medical intervention.
  - Prior chemotherapy or hormonal therapy for breast cancer. Patients treated with hormonal chemoprevention (tamoxifen or raloxifene) will be eligible.
  - Active malignancy diagnosed within the last 5 years. (Cervical cancer or non-melanomatous skin cancer that has been treated with curative intent will be eligible).

Results

- **Outcome Measurement:**
  - Number of Participants With Adverse Events.
  - This section provides the number of participants with adverse events. For a detailed list of adverse events, see the adverse event module.
  - Time frame: 6 years.
- **Results 1:**
  - **Arm/Group Title:** Dose A-Cohort 1-Arm 1-Docetaxel & Capecitabine.
  - **Arm/Group Description:** Docetaxel 75 mg/m^2 intravenous on Day 1, capecitabine 1000 mg/m^2 orally twice daily from Day 2 to Day 15 for 4 cycles. Once the dose was deemed to be too toxic, subsequent patients were enrolled on Dose B.
  - Overall Number of Participants Analyzed: 9
  - Measure Type: Number
  - Unit of Measure: Participants
    - Participants: 9
- **Results 2:**
  - **Arm/Group Title:** Dose B-Cohort 2-Arm 2 Reduced Dose-Docetaxel & Capecitabine.
  - **Arm/Group Description:** Docetaxel 60 mg/m^2 intravenous on Day 1, capecitabine 937.5 mg/m^2 orally twice daily from Day 2 to Day 15.
  - Overall Number of Participants Analyzed: 20
  - Measure Type: Number
  - Unit of Measure: Participants
    - Participants: 20

Adverse Events

- **Adverse Events 1:**
  - Total: 29/30 (96.67%)
  - Febrile neutropenia: 3/30 (10.00%)
  - Lymphatics: 1/30 (3.33%)
  - Diarrhea (without colostomy): 5/30 (16.67%)
  - Abdominal pain or cramping: 2/30 (6.67%)
  - Colitis: 1/30 (3.33%)
  - Dehydration: 1/30 (3.33%)
  - Nausea: 1/30 (3.33%)
  - Stomatitis/pharyngitis (oral/pharyngeal/mucositis): 1/30 (3.33%)
  - Vomiting: 1/30 (3.33%)
- **Adverse Events 2:**
  - (Data not provided)

Example 2

Clinical Trial ID

- Clinical Trial ID: `NCT00006110`

Interventions

- **INTERVENTION 1:** Herceptin Regimen After AC
  - Patients in the adjuvant and neoadjuvant groups after receiving [AC-TP] Chemotherapy (doxorubicin & cyclophosphamide).
- **INTERVENTION 2:** Herceptin Regimen After TP
  - Patients in the adjuvant and neoadjuvant groups after receiving chemotherapy and Taxol + Herceptin.

Eligibility

- **Inclusion Criteria:**
  - Histologically confirmed stage IIB, IIIA, IIIB, IIIC, or previously untreated stage IV primary carcinoma of the breast.
  - Fine needle aspiration, core needle biopsy, or incisional biopsy allowed.
  - No excisional biopsy.
  - Any of the following:
    * Tumor size 2 cm, Nodes 1 (T2N1) or tumor size 3 cm nodes 0 (T3N0).
    * Any T with N2 (including axillary lymph nodes matted to one another) or N3.
    * Any T4, including inflammatory breast cancer.
    * Adjuvant patients with at least 4 positive lymph nodes and HER-2 overexpressing tumor.

    * Supraclavicular or infraclavicular positive lymph nodes without distant metastases.
    * Distant metastases with measurable disease in breast or lymph nodes.
    * Synchronous bilateral primary breast cancer allowed if the more serious cancer meets entry criteria.
    * Measurable or evaluable disease.
  - **Patient Characteristics:**
    * Age: Not specified.
    * Sex: Female.
    * Menopausal status: Not specified.
    * Performance status: Not specified.
    * Life expectancy: Not specified.
    * Hematopoietic:
      · White cell count $> 3000\,\mathrm{mm}^{-3}$.
      · Platelet count $> 100{,}000\,\mathrm{mm}^{-3}$.
      · Hemoglobin $> 9\,\mathrm{mg/dL}$.
      · Bilirubin < 1.5 times normal.
      · Creatinine < 1.5 times normal.
      · Left ventricular ejection fraction (LVEF) normal by resting nuclear ventriculogram.
      · Not pregnant or nursing.
      · Negative pregnancy test.
      · Fertile patients must use effective contraception.

- **Exclusions:**
  - Prior malignancies except:
    * Effectively treated squamous cell or basal cell skin cancer.

    - Carcinoma in situ of the cervix that has been curatively treated by surgery alone.
    - Nonbreast malignancy from which patient has been disease-free for 5 years and is at low risk of recurrence.

Results

- **Outcome Measurement:**
  - Cardiac Toxicity of Weekly Taxol Given With Weekly Herceptin When Delivered Immediately Following Four Cycles of Standard Dose AC.
  - Doxorubicin + cyclophosphamide in combination with paclitaxel and trastuzumab (AC-TP) Associated Systolic Dysfunction. Systolic function was measured by the ventricular ejection fraction (LVEF). LVEF is a measurement in determining how well your heart is pumping out blood and in diagnosing and tracking heart failure.
  - Time frame: 78 weeks (1.5 years).
- **Results 1:**
  - **Arm/Group Title:** Herceptin Regimen After AC.
  - **Arm/Group Description:** Patients in the adjuvant and neoadjuvant groups after receiving [AC-TP] Chemotherapy (doxorubicin & cyclophosphamide).
  - Overall Number of Participants Analyzed: 52.
  - Measure Type: Count of Participants.
  - Unit of Measure: Participants
    - Asymptomatic LVEF $< 50\%$: 1 (1.9%).
    - Congestive Heart Failure: 0 (0.0%).
- **Results 2:**
  - **Arm/Group Title:** Herceptin Regimen After TP.

  - **Arm/Group Description:** Patients in the adjuvant and neoadjuvant groups after receiving chemotherapy and Taxol + Herceptin.
  - Overall Number of Participants Analyzed: 50.
  - Measure Type: Count of Participants.
  - Unit of Measure: Participants
    * Asymptomatic LVEF $< 50\%$: 8 (16.0%).
    * Congestive Heart Failure: 1 (2.0%).

Adverse Events

- **Adverse Events 1:**
  - Total: 7/52 (13.46%).
  - Febrile neutropenia (fever of unknown origin without clinically or microbiologically documented infection): 0/52 (0.00%).
  - Atrial Fibrillation: 1/52 (1.92%).
  - Sepsis: 1/52 (1.92%).
  - Muscle weakness upper limb: 1/52 (1.92%).
  - Dizziness: 1/52 (1.92%).
  - Seizure: 1/52 (1.92%).
  - Nervous system disorders - Other, specify: 1/52 (1.92%).
- **Adverse Events 2:**
  - Total: 1/30 (3.33%).
  - Febrile neutropenia (fever of unknown origin without clinically or microbiologically documented infection): 1/30 (3.33%).
  - Atrial Fibrillation: 0/30 (0.00%).
  - Sepsis: 0/30 (0.00%).
  - Muscle weakness upper limb: 0/30 (0.00%).

- Dizziness: 0/30 (0.00%).
- Seizure: 0/30 (0.00%).
- Nervous system disorders - Other, specify: 0/30 (0.00%).