



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Thesis for the Degree of Doctor of Philosophy

**KNOWLEDGE EXTRACTION FROM
UNSTRUCTURED CLINICAL TEXT USING
ACTIVE TRANSFER LEARNING APPROACH**

Musarrat Hussain

**Department of Computer Science and Engineering
Graduate School
Kyung Hee University
Yongin, Korea**

August 2022

KNOWLEDGE EXTRACTION FROM UNSTRUCTURED CLINICAL TEXT USING ACTIVE TRANSFER LEARNING APPROACH

Musarrat Hussain

**Department of Computer Science and Engineering
Graduate School
Kyung Hee University
Yongin, Korea**

August 2022

KNOWLEDGE EXTRACTION FROM UNSTRUCTURED CLINICAL TEXT USING ACTIVE TRANSFER LEARNING APPROACH

by

Musarrat Hussain

Supervised by

Prof. Sungyoung Lee

Prof. TaeChoong Chung

Submitted to the Department of Computer Science and Engineering and the
Faculty of Graduate School of Kyung Hee University in partial fulfillment of the
requirements of the degree of Doctor of Philosophy

Dissertation Committee:

Prof. Sung-Ho Bae


Prof. Seok-Won Lee

Prof. Seungkyu Lee

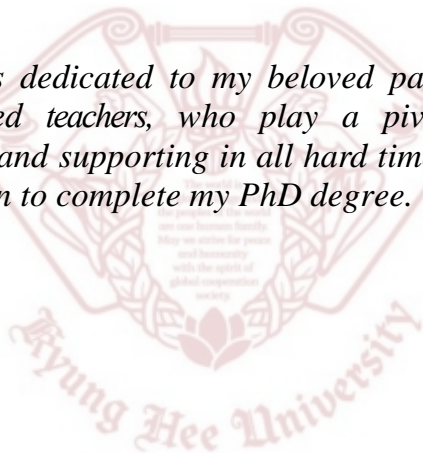
Prof. Seong Tae Kim

Prof. TaeChoong Chung

Prof. Sungyoung Lee



*This thesis is dedicated to my beloved parents, family,
and respected teachers, who play a pivotal role by
encouraging and supporting in all hard times to keep me
in the position to complete my PhD degree.*



Abstract

The availability of healthcare data and advancements in computing technologies have enabled access to knowledge from a plethora of sources. Artificial intelligence (AI) plays an important role in knowledge acquisition and intelligent decision-making. The AI-based systems evaluate annotated patient data to identify and learn the human experts' decision-making heuristics. The intelligence of these systems is primarily dependent on the data used for knowledge acquisition. Huge data covering diverse aspects leads to more accurate knowledge acquisition. However, most of the clinical data is stored in unstructured textual format which cannot be directly processed by machines for knowledge extraction. Natural language processing (NLP) provides a way to extract meaningful information from texts, which is a tedious task for human beings to process the plethora of textual documents. The existing NLP applications for healthcare services focus on individual tasks such as text classification, clinical concept extraction, or relation extraction which have achieved great success in the individual task. However, there is an utmost need for an end-to-end methodology to extract executable knowledge from the plethora of clinical textual data so that the acquired knowledge can assist clinical decisions and improve healthcare quality.

This dissertation proposes and investigates the creation, application, and evolution of a knowledge extraction methodology in the clinical domain. Using state-of-the-art modular solutions, a novel knowledge processing pipeline transforms unstructured clinical text into production rules. First, input clinical text is cleaned and classified to produce a set of recommendation and non-recommendation sentences. This classification is performed using a combination of statically defined, expert-sourced, and probabilistically determined, data-sourced patterns. Next, the set of recommendation sentences is processed to extract clinical concepts and identify the syntactic relationship between them. The disjoint union of concepts and their inter-relationships forms

knowledge triples which act as the atomic source for further processing. Using active transfer learning, this knowledge base is built and evolved to accurately identify clinical causes and their effects. Finally, triples are transformed into production rules to support transparency and an expert validation of the knowledge bases.

The combination of expert-driven and data-driven patterns in the first module of the proposed pipeline achieved an accuracy of 76.92% for asthma, 85.32% for rhinosinusitis, and 92.07% for hypertension guidelines. The experimental results show our text classification module significantly outperforms the expert-driven pattern by 1.97%, 0.69%, 2.91%, and data-driven (Decision tree) patterns by -4.8%, 2.93%, and 8.38% on Asthma, Rhinosinusitis, and Hypertension guidelines, respectively. In the next step, the SemEval training dataset was used for causal triple learning, its test part for threshold selection, Asian Bayesian network, and Alzheimer's disease (AD) datasets for evolving the causal identification model. The application of active learning on the set of causal sentences in the newly built model improved its performance in terms of F1 score, which increased by a factor of six percentage points after the first iteration and one percentage point after the second iteration. In comparison to the previously existing solutions, this methodology outperformed by 1.11 percentage points. Finally, the evaluation of value extraction for causal concepts achieved an accuracy of 91.55% for the hypertension guidelines. The proposed knowledge extraction pipeline is able to achieve the aforementioned objectives, thereby providing a sound realization of unstructured data to knowledge transformation.

Acknowledgement

Alhumdolillah, By grace of ALLAH Almighty, who is the most beneficent and merciful. Who gave me the strength, courage, patience during my doctoral study and showering HIS blessings upon me and my family.

I am highly grateful to my advisors Prof. Sungyoung Lee and Prof. Tae Choong Chung for their boundless moral and technical supervision, guidance, and courage in coping with the difficult challenges throughout the education period of my doctoral studies. They trained me in multi-directions to face the challenges of practical life in a professional manner. Their lively natures, clear assistance, and direction enabled me able to complete my thesis. They have refined the key ingredients for high quality research, namely my skills of creativity, thinking, and technical understanding. Moreover, I would like to acknowledge their valuable guidance and support to refine the problem statement as well as that to streamline my research direction.

I appreciate my dissertation evaluation committee for their valuable observations and insight recommendations during the dissertation defense. These comments enhanced the presentation and contents of the dissertation.

I am extremely grateful to all of my current and former Ubiquitous Computing Lab fellows and colleagues who have always provided me time, expertise, and encouragement. They were always present to guide me in various situations throughout my PhD journey. I would like to thank Dr. wajahat Ali Khan, Dr. Bilal Amin, Dr. Maqbool Hussain, Dr. Muhammad Afzal, Dr. Taqdir Ali, Dr. Jamil Hussain, Dr. Maqbool Ali, Dr. Jae Hun Bang, Dr. Tae Ho Hur, Dr. Shujaat Hussain, Dr. Muhammad Asif Razzaq, Dr. Usman Akhtar, Dr. Bilal Ali, Dr. Mugahed A. Al-antari, Dr. Huacam Hao, and Mrs. Seoungae Kim. They have contributed enormously in successfully performing various academic and personal tasks that confronted me during my stay at South Korea.

I am very thankful to all of my colleagues for their kind support to my personal and academic life at Kyung Hee University. I am highly obliged to brilliant researchers Syed Imran Ali, Fahad Ahmed Satti, Ubaid Ur Rehman, Muhammad Sadiq, Anees ul Hassan, Asim Abbas, Muhammad Zaki Ansaar, Dr. Waseem Hassan, Dr. Dildar Hussain, Dr. Saeed Ullah, Dr. Ahsan Raza Kazmi, Dr. Ibrar Yaqoob, Zain Abbas, and Ahsan Raza. This journey would not have been possible without their support. They contributed a lot in to my personal and academic life to polish myself. Also, I appreciate all my Korean and international friends who worked as a team with me and developed my team work skills and provided me wonderful memories during my stay in South Korea. I would like to extend my sincere thanks to my kind and respected teachers at primary, high school, college and university level, who support and encourage me to pursue my higher study.

Last but not the least, I would like to express my sincere gratitude to my parents, sisters, and brothers for their endless love, support, prayers, and encouragement. Their support and encouragement has made this dissertation possible. Particularly, my mother (may ALLAH rest her soul in peace), who keep supporting to continue my study. Although, she did not see the completion of my PhD study, however, it was her dream that ultimately come true. I would like to extend my thanks to my wife, friends in Pakistan (especially Mr. Muhammad Ejaz), and other relatives who provide their kind support and encouragement to follow my dreams

Musarrat Hussain

August, 2022

Table of Contents

Abstract	i
Table of Contents	v
List of Figures	viii
List of Tables	xi
Chapter 1 Introduction	1
1.1 Motivation	1
1.2 Problem Statement	4
1.3 Proposed Methodology Overview	5
1.4 Key Contributions	8
1.4.1 Machine Learning Assisted Pattern Extraction for Text Classification	8
1.4.2 Semantic Pattern Extraction Algorithm	9
1.4.3 Features Expansion of ML based Classification	9
1.4.4 Phrase based Causality Mining	10
1.4.5 BERT based Multi-Model Approach	10
1.4.6 Active Learning Approach	10
1.5 Thesis Organization	11
Chapter 2 Related Work	13
2.1 Clinical Text Mining	13
2.2 Literature Summary	14

2.2.1	Clinical Text Classification	14
2.2.2	Causality Mining Related Work	16
2.2.3	Clinical Concepts' Value Extraction	18
Chapter 3	Proposed Methodology	21
3.1	Clinical Text Classification	23
3.2	Causality Mining	25
3.3	Rules Generation	27
Chapter 4	Clinical Text Classification	29
4.1	Pattern based Classification	29
4.1.1	Heuristic Pattern Extraction	30
4.1.2	Machine Learning Assisted Patter Extraction	31
4.1.3	Automatic Pattern Extraction	38
4.2	Traditional Machine Learning based Classification	40
4.3	Deep Learning based Classification	42
Chapter 5	Causality Mining	44
5.1	Preprocessing Module	45
5.2	Model Development (MD) Module	47
5.2.1	Causal Trigger Extraction	47
5.2.2	Model Training/Evolution	48
5.2.3	Example of Model Development (Training Phase)	48
5.3	Causality Mining (CM) Module	49
5.3.1	Candidate Triple Extraction	49
5.3.2	Causal Candidate Classification	50
5.3.3	Triple Semantic Analyzer	51
5.4	Feedback Loop	51
Chapter 6	Rules Generation	54
6.1	Concepts Operator and Value Identification	55

6.2	Example of Concepts Operator and Value Identification	56
6.3	An End-to-End Example	58
Chapter 7	Results and Evaluation	61
7.1	Text Classification Results	61
7.1.1	Text Classification Dataset	61
7.1.2	Pattern based Classification Results	62
7.1.3	Traditional Machine Learning based Classification Results	68
7.1.4	Deep Learning based Classification Results	69
7.2	Causality Mining Results	71
7.2.1	Experimental Setup	71
7.2.2	Stage 2 - Threshold Selection Results	75
7.2.3	Experimental result with Word2Vec embeddings	77
7.2.4	Experimental Result with BERT Embeddings	78
7.2.5	Experimental Result with BioBERT Embeddings	80
7.2.6	Single Model Evaluation	81
7.2.7	Multi-model Evaluation	84
7.2.8	The Feedback Loop Evaluation	85
7.2.9	Comparison with existing studies	86
7.2.10	Discussion	87
7.3	Rules Generation Results	90
7.3.1	Experimental Setup	90
7.3.2	Concepts Values Extraction Results	91
7.4	End-to-end Evaluation	92
Chapter 8	Conclusion and Future Direction	94
8.1	Conclusion	94
8.2	Future Direction	96
	Bibliography	97
	Appendix A List of Acronyms	108

Appendix B	List of Publications	110
B.1	International Journal Papers [9]	110
B.2	Domestic Journal Papers [4]	111
B.3	International Conference Papers [10]	112
B.4	Patents [4]	113



List of Figures

1.1	Types of knowledge required for making clinical decisions	3
1.2	An overview of the proposed methodology	6
3.1	The proposed methodology workflow	22
3.2	Text classification approaches with pros and cons	23
3.3	Causality mining abstract view	26
3.4	Rules generation abstract view	28
4.1	The proposed pattern extraction methodology for clinical text classification.	30
4.2	Example decision tree model for salient term extraction.	33
4.3	Example of UMLS based pattern extraction	37
4.4	Semantic pattern extraction example	40
4.5	Machine learning pipeline for text classification	42
5.1	Causality mining workflow	45
5.2	Training causal trigger extraction example	49
5.3	Training causal trigger extraction example	50
6.1	Rules generation workflow	54
6.2	Example of rules generation from causal triples	57
6.3	Example sentence parsed via the Stanford parser	58
6.4	An end-to-end example of the proposed methodology	59
7.1	Top k features and the model accuracy	63
7.2	Models accuracy without and with features selection	63

7.3	Extracted patterns accuracy	64
7.4	Combined patterns accuracy with and without salient terms.	65
7.5	Extracted patterns evaluation	67
7.6	Machine Learning Models Classification Results	68
7.7	Evaluation of proposed method on large datasets	69
7.8	Causality mining experimental setups	71
7.9	Details of causality mining dataset.	73
7.10	Precision recall curve for threshold selection (a) bert-base-nli-mean-tokens (b) bert-base-nli-max-tokens (c) bert-base-nli-cls-tokens (d) bert-large-nli-mean-tokens (e) bert-large-nli-max-tokens (f) bert-large-nli-cls-tokens.	76
7.11	Precision recall curve for threshold selection for BioBert.	81
7.12	UpSet analysis of BERT model classification coverage for a combined list of Risk Factors of Alzheimer's Disease and Asia Bayesian Network dataset.	84
7.13	End-to-end methodology evaluation.	93

List of Tables

2.1	Summary of text classification approaches	16
2.2	Summary of causality mining approaches	18
4.1	Evaluation matrix for nominal group technique (NGT).	32
4.2	Extracted heuristics patterns	32
4.3	List of salient terms considered by machine learning models	34
4.4	Extracted heuristics patterns with salient terms	35
4.5	List of used POS tags	36
4.6	List of extracted POS patterns	36
4.7	List of extracted UMLS Patterns	37
7.1	Details of text classification dataset	61
7.2	Initial Experiments with Word2Vec based embedding vector generation on SemEval Test dataset	78
7.3	Setting 2 with BERT based embedding vector generation on SemEval Test dataset	79
7.4	Application of trained embedding on Asia Bayesian Network dataset	80
7.5	Application of trained embedding on Risk Factors of Alzheimer's Disease dataset	80
7.6	Application of BioBert Embedding on Test Datasets	81
7.7	Application of trained embedding on Asia Bayesian Network dataset	
	Legend: TP is True positive, FN is False Negative, FP is False Positive, TN is True Negative, A is accuracy, P is precision, R is recall, and F1 is F1 Score	82
7.8	Application of trained embedding on Risk Factors of Alzheimer's Disease Split 1	83
7.9	Application of trained embedding on Risk Factors of Alzheimer's Disease Split 2	83

7.10 Application of Multimodel Embedding on Test Datasets	85
7.11 Feedback loop results on test datasets	86
7.12 Result comparison with Ning's method on test datasets	87
7.13 Concepts value extraction result	91
7.14 End-to-end methodology results	92



Technological advancements have greatly influenced the healthcare industry by enhancing its reach to a wider population pool and augmenting clinical practices with state-of-the-art research. The advanced Artificial Intelligence (AI) technology of the current era is working alongside human experts in assisting clinical decision-making and improving healthcare deliveries [1]. The intelligence of AI systems is primarily acquired from data of the domain. The amount of healthcare data is tremendously increasing over time [2]. Approximately 80% of healthcare data is stored in machine non-understandable (unstructured) format. Manual processing of this valuable clinical data is a challenging and resource-intensive task, especially in the time-critical clinical scenario [3]. Therefore, this dissertation investigates an automatic clinical text processing methodology for knowledge acquisition. In particular, this research investigates three essential aspects of knowledge acquisition, including distinguishing valuable content from the background information via text classification, clinical concepts and their relation extraction using causality mining technique, and executable knowledge creation from the extracted concept with their relations. In this way, a significant amount of physicians' time and burden can be reduced and the acquired knowledge can improve the intelligence of AI systems in assisting clinical decisions as well as human experts for providing informed decisions.

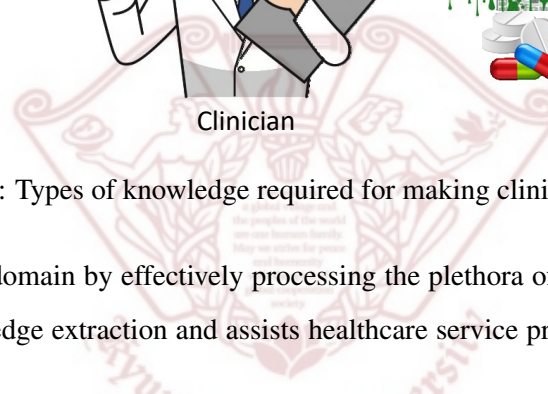
1.1 Motivation

Healthcare service providers required various background knowledge for making clinical decisions [4]. Among all, three types of knowledge are compulsory, including general medical knowledge, patient profile information, and medical procedures [5, 6]. As shown in Figure 1.1, general medical knowledge deals with disease information, their symptoms, and available medications

for treating these diseases. Profile data provides patient-related information such as disease history, allergy, family history, and others which helps in the diagnosis, treatment, and prescribing follow-up plans. While medical procedures define what action and procedure should be followed in various conditions. Medical procedures are described in various clinical documents such as Clinical Practice Guidelines (CPG), previous practices reported in clinical notes, clinical protocols, online research articles, usually written in unstructured textual format. The number of these documents is increasing at an enormous rate [7]. Over 35,000 CPGs are indexed by PubMed¹ with over 1500 appearing every year [8]. Similarly, the number of clinical notes and clinical findings are approaching beyond human capacity. These documents have various challenges, namely grammatical errors, spelling mistakes, use of domain-specific abbreviations, and negation expressions, among others [9]. The free text format of these documents and in-built challenges make it difficult to process, resulting in the under-utilization of these valuable knowledge sources [10]. Thus, the latest knowledge remains limited to documents without affecting real practices leading to a huge gap between research findings and real practices. This gap needs to be eliminated by the latest AI technology assistance to reduce physicians' burden, mitigate processing costs, save their time, and improve healthcare services by providing the best possible practices.

Digital evolution in the last century has greatly accelerated innovations and accomplished remarkable achievements in numerous domains. In particular, the healthcare domain has attracted a multitude of researchers to supplement AI base solutions for providing the right services at right time [11]. The precious information buried in the healthcare system as free text exhibits itself as a leading contender for Natural Language Processing (NLP) applications. NLP aims to effectively complete tasks concerning natural language by enabling machines to understand and process written languages in words, sentences, or paragraphs [12]. Early NLP applications were mostly inclined towards machine translation applications to translate text from one language to another through direct content mapping [13]. The dictionary and hand-crafted rule-based NLP processing were enhanced with machine learning (ML) to better understand and process the natural language text. The latest AI based NLP solutions such as BERT [14] has already revolutionized natural language processing by providing state-of-the-art applications. This revolution needs to

¹<https://pubmed.ncbi.nlm.nih.gov/>



domain by effectively processing the plethora of
knowledge extraction and assists healthcare service pr

edge extraction and assists healthcare service pr

ts of AI technologies and the increasing number of specialized NLP-based solutions for automatic processing of meaningful information and knowledge extraction for processing the challenging nature of clinical knowledge acquisition. The acquired knowledge can be used as in automated systems such as Clinical Decision Support Systems for clinical decisions. Thus, it bridges the gap between clinical documents and real practices. The proposed solution reduces the time, effort, and cost of the clinical experts.

1.2 Problem Statement

Clinical text such as clinical practice guidelines, clinical notes, discharge summaries, etc. withholds implicit knowledge [15], which provides a rich source for applying and enhancing clinical practices [16, 17]. These documents describe disease-specific process flows, patients' summaries, medical decisions, and medication details, which provide the necessary ingredients for dealing with a wide variety of medical situations [18, 19]. However, knowledge acquisition remains a problem due to the complex nature of these documents. It aggregates background information along with disease-specific information. Also, the extraction of clinical entities and their relationships is an open problem despite enormous research efforts [20]. Identification of a machine-readable representation of this knowledge necessitates a stable, scalable, and semi-automatic mechanism. To design an effective text mining methodology for machine-understandable knowledge acquisition, the following research questions must be answered:

1. How to distinguish between recommendation and non-recommendation content in the clinical text? Clinical text classification is an essential step of clinical text mining with a significant effect on subsequent steps. It aims to identify patient-related information, symptoms, lab tests, diagnosis, treatments, follow-up, and clinical procedures relevant text while filtering out irrelevant contents. Clinical text classification is an exception and challenging task because it requires a large amount of domain knowledge [21]. Erroneous classification of recommendation contents results in incomplete knowledge acquisition while classifying non-recommendation text as recommendation leads to incorrect knowledge acquisition and ultimately inappropriate decisions. Therefore, this dissertation explored diverse techniques, including pattern-based and machine learning-based approaches to address clinical text classification challenges efficiently.
2. How to identify clinical concepts and their relationships in the text? The clinical concepts identification known as named entity recognition (NER) task seeks to locate clinical terminologies and assign it an associated category such as sign symptom, disease disorder, severity, diagnostic procedure, medication, medication dosage, etc. [22]. While relation extraction aims to determine the association between entities. The relation understanding

among various entities is significantly important as it helps in various clinical tasks such as disease cause identification, drug-drug interaction, patient response to drug identification, and others [23]. Among all possible relations, cause-effect represents an essential relation, which provides ample support for the reasoning and decision-making process for humans as well machines [24]. Therefore, this dissertation addressed clinical entities and relation extraction challenges, with particular emphasis on the cause-effect for better understanding textual content, assisting knowledge acquisition, and providing informed quality decisions.

3. How to determine the status and mentioned values of the clinical entities such as symptoms, lab test results, diseases, treatments, medication-related information, etc.? The clinical text is composed of patient-associated conditions and other related measures such as gender, age, weight, height, vital signs, and others [25]. Extracting values for these clinical entities is a challenging and critically important task as it provides the ground for making a diagnosis, treatment, and follow-up decisions. This dissertation investigates traditional (pattern-based approaches) as well as state-of-the-art machine learning approaches for attribute values extraction. The extracted attributes, their relationships, and values can be represented in a machine-understandable format such as production rules to be used by decision assistive systems for making appropriate decisions and improving healthcare quality.

1.3 Proposed Methodology Overview

In this work, we proposed a methodology for clinical text mining that transforms the input text into both human and machine understanding knowledge. The methodology comprises of three solutions; sentence classification, causality mining, and rule generation, as depicted in Figure 1.2. The primary objective of Solution 1: Sentence Classification is to classify clinical sentences into recommendation and non-recommendation sentences based on the importance of the presented contents. Solution 2:Causality Mining extracts clinical concepts and identifies their relationships. In the clinical domain, healthcare experts are mainly interested in casual relationships. Therefore, this dissertation focuses on causality mining for relationship identification. While Solution 3: Rule Generation aims to represent the identified concepts and relationships in production rules format

to fulfill the final goal of human and machine-understandable knowledge acquisition.

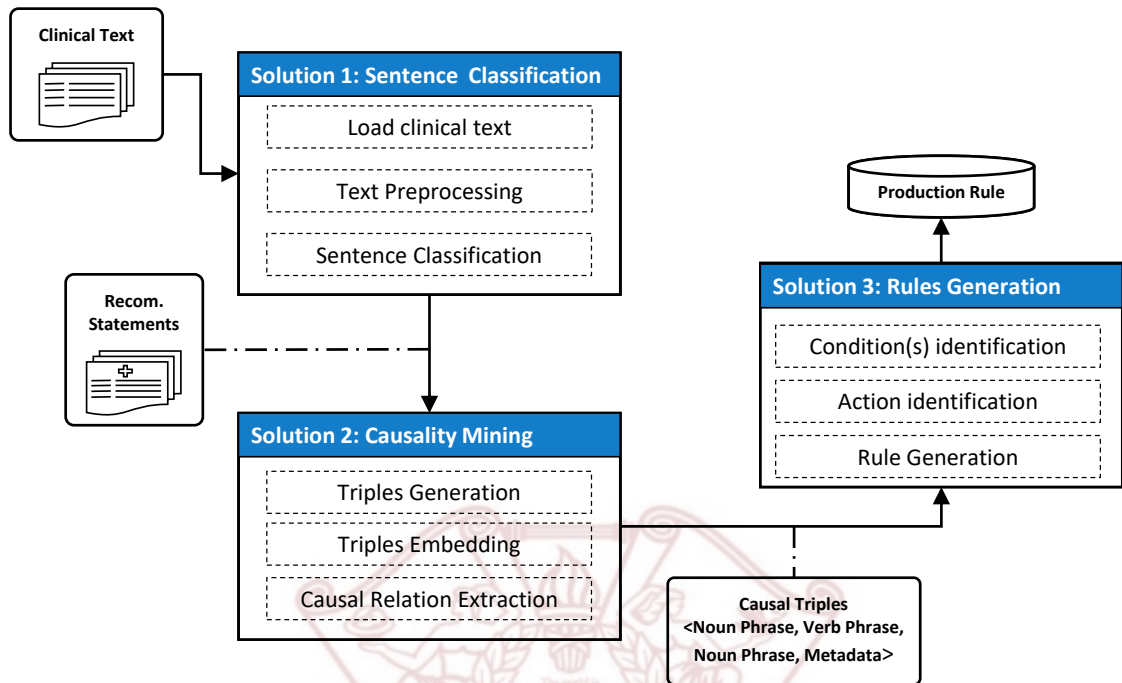


Figure 1.2: An overview of the proposed methodology

Based on the importance of the provided information, Clinical contents can be categorized into two parts. First, the background information, which includes abstract information related to the background and point of view of the authors. Second, the disease-specific information, which elaborates causes, consequences, and actions related to a disease. Therefore, the understandability and classification of clinical contents is an important step, before its transformation to computer interpretable format. Among this information, the recommendation sentences are the main focused and desired contents that need to be extracted from clinical text. These content assist the domain experts in making evidence based decisions.

The field of text classification and information extraction has greatly benefited from advances in computing, producing a plethora of algorithms, tools, and applications, based on machine-learning and pattern-based approaches [26–31]. However, in the clinical domain, most of the natural language processing tasks including, guideline processing and information extraction, are still using pattern-based approaches [32]. Pattern-based approaches perform better than machine-

learning models in clinical text classification, because of the involvement of human experts for pattern extraction [33]. To reduce the manual effort and reflect machine intelligence for text classification, this research focuses on machine learning assistive pattern extraction and their generalization. Where human experts are assisted by machine learning methods during the pattern extraction process to come up with the most appropriate patterns. Also, an automatic pattern extraction algorithm has been proposed that extracts semantic patterns for efficient clinical text classification. The details of the text classification methodology can be found in Chapter 4 of the dissertation.

The recommendation tagged sentences of the clinical text are further processed for clinical concepts and relationship identification. There exist various relationships between concepts including, “improves”, “worsens”, “given”, “reveals”, and others [34]. However, we primarily focused on the cause-effect relation between concepts. A combination of transfer learning and active learning (active transfer learning) mechanism is adopted for causality mining. Causality detection is typically based on two tasks, the identification of causal triggers also known as causal connectives, and causal pairs participating in each relationship [35]. Therefore, causality mining techniques follow the Noun Phrase (NP) - Verb (V) - NP pattern which corresponds to either Cause - Trigger - Effect or Effect - Trigger - Cause forms ($\langle S \rightarrow NP\text{-Cause}, V \rightarrow \text{Verb-Trigger}, O \rightarrow NP\text{-Effect} \rangle$) [36]. In this research, we first identifies causal phrases in the form of causal triples (subject, causal verb, and object) using dependency based linguistic patterns from a training dataset. Each component of the causal triple is then expanded via transfer learning using pre-trained Google News model [37]. The expanded causal triple in “NP V NP” (SVO) form is then converted into embedded vector using Bidirectional Encoder Representations from Transformers (BERT) [38]. These embedded vectors are then used to calculate a similarity matching score, against embedded causal triples from unseen test dataset. The matching scores, and evaluation of the precision-recall curve then provides the matching threshold, over which a triple (and its corresponding phrase) can be classified as causal and under which as non-causal. The embedded vectors from the training dataset and the threshold calculated thus far, are then applied on two test datasets, to classify each test triple as causal or non-casual. The noun phrases within these causal quads (extended triple with similarity score) are then semantically enriched using Unified Medical

Language System (UMLS). The resultant quads are then validated by expert and used for causal triple extension using active learning methodology. The details of the proposed causality mining methodology are described in Chapter 5.

The causal classified triples are evaluated for rules generation. As production rules follow IF condition(s) THEN conclusion pattern, where the condition consists of Key, Operator, and Value. We mapped the causal concept of a triple into condition, while the effect to the conclusion part of the rule. For the condition part, we also need the operator and a value therefore, we evaluated the neighbor tokens of the causal concept for possible value extraction. A nearest “quantitative” token of a casual concept is set as the value of the concept and “Quantitative Concept” for the operator of the conditional concepts, which completes the rule condition. Conditions with similar effects are combined to build the final set of rules. The acquired rules are understandable to humans as well as can be used by decision systems for supporting clinical decisions. The details of the proposed rule generation are described in Chapter 6.

1.4 Key Contributions

The goal of this research work is to provide an end-to-end methodology of clinical text mining for human and machine-understandable knowledge acquisition. The proposed methodology is devised with three main objectives of clinical text classification, clinical concepts and relation extraction, and rule generation. We achieved the mentioned objectives with the following research contributions.

1.4.1 Machine Learning Assisted Pattern Extraction for Text Classification

Text classification is an essential step of clinical text mining. Despite the advancement of machine learning techniques, pattern-based classification is still a preferable and widely used approach for clinical text classification. Traditionally, domain experts manually analyze the input text for patterns extraction, which is a tedious and time-consuming task. Also, the resultant patterns are dependent on the intuition and expertise of human experts. This research proposed a machine learning-based pattern extraction methodology to reduce the human experts’ burden and augment

machine intelligence with human intelligence for patterns generation. Machine learning methods extract salient terms contributing to distinguishing recommendation and non-recommendation sentences. Human experts incorporate the extracted salient terms with their experience and heuristics to produce more effective patterns. Also, we generalize the extract patterns to POS and UMLS patterns, which generalize the proposed method and make it applicable to other clinical texts.

1.4.2 Semantic Pattern Extraction Algorithm

As described, pattern extraction approaches commonly require human intervention for pattern identification, which diminishes their benefits and restrain their applications. Also, the extracted patterns depend on specific clinical terms and suffer from the specialization problem due to synonymy and polysemy. In this research, we proposed a novel pattern extraction algorithm, which identifies and extracts patterns from clinical textual resources, automatically. The algorithm identifies the candidate clinical terms in the text, determines their semantic types from UMLS, finds the context of the concepts by discovering their context windows, and finally transforms each context window into a pattern. The extracted semantic patterns mitigate patterns specialization problems and increase the scope and performance of the resultant patterns.

1.4.3 Features Expansion of ML based Classification

Machine learning based text classification transforms the input text into a structured format, applies various models such as Naive Bayes, Decision Tree, Random Forest, etc. to learn insights. The trained model then can classify similar unseen clinical text. The learning of ML models are critical depended on the feature extracted from the training text. Therefore, in this research we explored the effect of features expansion on the performance of various ML Models. The features were enriched by features synonyms from WordNet [39] and semantic categories from UMLS. As, the expanded features provide more broader and meaningful information therefore, the ML models were able to better understand and learn the classification logic. As a result, the expanded features increases the ML models classification performance.

1.4.4 Phrase based Causality Mining

Causality mining aims to determine whether two entities are having cause-effect relationships or vice versa, from an input text. Typically, the relationship is decided solely on a verb term also known as a causal trigger or causal connectives, connecting the two entities. This research investigates a phrase/triple-based causality mining encapsulating causal triggers with target entities to preserve their semantics. Thus, we generate casual triples of the form $\langle \text{NP}, \text{VP}, \text{NP} \rangle$ from training data and expand the participating terms using word expansion mechanism to increase the scope and converge of the training triples. Similarly, the unseen test data is processed to generate triples. The proposed phrase/triple-based approach maintains the semantics of the target terms and increases causal relationship identification.

1.4.5 BERT based Multi-Model Approach

The objective of transfer learning is to transfer knowledge learned in a domain into a new relevant domain. BERT is the state-of-the-art language model trained on English text for improving NLP tasks. This research utilized the BERT learned knowledge to encode the extracted triple phrase into meaningful vectors. The trained embedded vectors are then used to compare with test embeddings for similarity evaluation and casual relation mining. However, we discovered by analyzing various BERT model results that different models can detect various unique casual phrases. Therefore, we combined six BERT models namely nli-base-mean-tokens, nli-large-mean-tokens, nli-base-max-tokens, nli-large-max-tokens, nli-base-cls-token, and nli-base-cls-token to form a multi-model approach. A triple/phrase tagged causal by any model was considered causal, which increased the overall performance of causality mining.

1.4.6 Active Learning Approach

One of the main drawbacks of ML models is performance degradation over time. These models are not able to maintain their knowledge up-to-date by incorporating experts' feedback and new knowledge. This research incorporates expert feedback to evolve the underlying training set of the model over various runs incorporating active learning methodology. An expert verifies the newly extracted causal phrases/triples for their correctness. The correct identified phrase/triples

are added to the causal training triples and the incorrectly identified triples are used to refine the training model by removing similar triples from the training set. This active learning process increases individual BERT model as well as Multi-model approach performance over various runs.

1.5 Thesis Organization

This dissertation is organized into chapters as follows.

- **Chapter 1: Introduction.** Chapter 1 deals with the overview of clinical text processing for knowledge acquisition. In this regard, the motivation for natural language processing based knowledge acquisition is emphasized. Furthermore, problem statement and overview of the proposed methodology is also put forward in this chapter. In the end, the key contributions of the dissertation are discussed.
- **Chapter 2: Related Work.** Chapter 2 focuses on the literature review for similar approaches for clinical text classification, causality mining, and rule generation. The key limitations of the existing approaches are also identified and enlisted here. Finally, this chapter summarizes how the identified limitations are mitigated via the proposed solutions.
- **Chapter 3: Proposed Methodology.** In Chapter 3, we presented the proposed end-to-end methodology for clinical knowledge acquisition from unstructured clinical text. This chapter deals with the three building blocks of the methodology, namely, clinical text classification, causality mining, and rules generation.
- **Chapter 4: Clinical Text Classification.** Chapter 4 provides the detail of various efforts made for robust classification. It elaborates the proposed ML assisted pattern extraction process. Additionally, it describes the proposed automatic semantic pattern extraction algorithm.
- **Chapter 5: Causality Mining.** Clinical concepts and their relationships play a key role in knowledge acquisition. As stated, this dissertation focuses only on the cause-effect relation between entities. Chapter 5 describes the detail process of entity extraction, their expan-

sion, phrase generation, BERT based phrase embedding and semantic matching, semantic enrichment, expert verification, and model evolution.

- **Chapter 6: Rules Generation.** Chapter 6 provides detail about processing the cause-effect phrase to production rule generation. It provides a comprehensive method of value identification for causal concepts of the cause-effect phrases. Also, the rule generation from the causal concepts with the identified values and effect concepts are explained in this chapter.
- **Chapter 7: Results and Evaluation.** The results and evaluation of various modules of the proposed methodology are explained in Chapter 7. Firstly, it explains the text classification result for the pattern and ML approaches. Secondly, the causality mining results are described for the individual six BERT Models. The result analysis and comparison with the existing approach are also explained. Finally, the concepts' value extraction and rules generation results are described.
- **Chapter 8: Conclusion and Future Direction.** This Chapter 8 concludes the thesis and also provides future directions in this research area. The main contribution of the thesis is also highlighted in this chapter.

Patient clinical histories, such as previously diagnosed diseases, provided treatments, allergies, and others, play a pivotal role in healthcare decisions. The requirement of clinical record tracking and timely access initiate the idea of clinical record storage and maintenance systems such as EHRs. The history of clinical records can be linked back to the fifty century B.C. when Hippocrates specified two of its aims, including correctly reflecting the course and potential cause of a disease [40]. The modern EHRs started to appear in the 1960s, supplementing the prescribed goals with additional functionalities. The clinical records contain both structured and unstructured information however, about 80% of clinical observations are not directly machine-understandable due to its unstructured format [41]. The unstructured clinical text is one of the most significant barriers of EHRs and clinical data in quality improvement, operations, and clinical research [42]. Starting from the 1940s till to date, Natural Language Processing (NLP) has made tremendous advancements in processing narrative text for various tasks such as Machine Translation, Automatic Summarization, Co-Reference Resolution, Discourse Analysis, Named Entity Recognition, information extraction, etc. [43, 44]. The NLP-based solutions for clinical text processing are summarized in the upcoming sections.

2.1 Clinical Text Mining

Clinical text withholds valuable information, including symptoms, diagnosis, treatment, medication details, and follow-up plans that can help in improving healthcare service provision. Clinical text mining refers to the automatic processing of a clinical text for understanding and interpretation of the content [9]. Plenty of research has been conducted to extract valuable information out of this text. The field of clinical text mining has advanced rapidly transitioning from hand

crafted rule based methods to machine learning and recently more advance approaches such as deep learning for information extraction and modeling [42].

2.2 Literature Summary

As discussed, this dissertation divides clinical text mining into three main tasks including text classification, clinical concepts and relation extraction (causality mining), and concepts values extraction. The existing approaches for each task are described as follows.

2.2.1 Clinical Text Classification

Clinical text classification is one of the essential and widely studied steps for clinical text mining [16]. The classification approaches can be divided into two main categories including pattern based approaches, and machine learning based approaches. Kaiser et al. [45] proposed a pattern based approach for detecting action and procedures in clinical text (CPG). The authors used UMLS classes to identify patterns which employed for activities representation and the semantic relations among them. The study consists of four steps. In the first step, they analyzed CPG regarding actions and procedures. In the second step, they explored the relationship between actions and procedures. In the third step, they expand the semantic type of the identified relation for generalization. Finally, they generated a dictionary of the identified actions, procedures, and their relations. The identified patterns in the dictionary were used to distinguish action, procedure and background related text. R Wenzina et al. [46] proposed a rule-based method using a combination of linguistic and semantic information of UMLS semantic type. The authors hypothesized that each guideline statement had its owns domain dependent linguistic and semantic patterns. They also induce weighting coefficient called relevance rate that shows statements relevancy for modeling. The relevance rate enables the authors to identify the condition-action combination. The relevance rate shows that the statement is crucial for a clinical pathway. Ashtma guideline was used for pattern extraction. The patterns extracted from the guideline were consisted of 12 “if” and 4 “should” statements. The analysis showed that rules of type “if” has a better result than the one of type “should”.

H. Hematialam et al. [47] designed supervised learning models including ZeroR, Naive Bayes, J48, and Random Forest for the classification of CPG statements. The model classifies a CPG statement into no-condition (NC), condition-action (CA), or condition-consequence (CC). The models were trained on three annotated guidelines (Hypertension, Chapter 4 of asthma, and rhinosinusitis) using Part of Speech (POS) as a feature to remove domain dependency constraints. The recommendation statements were identified by using modifiers and regular expressions. The most commonly used modifiers were "if", "in", "to", "for", "when", and "which". The identified recommendation statements were transformed to the "if condition then consequences" format for rule generation in later stages. The authors' used models were one shot models and required retraining each time when a change occurs in the training dataset.

S. Priyanta et al. [48] performed a comparative analysis of sentence subject classification using rule-based and machine learning models. The authors used opinion patterns for rule generations. The sentence subjectivity evaluation was performed on Indonesian news to classify a news sentence into subject or objective. The machine learning models Naive based classifier (NBC) and multinomial Support Vector Machine (SVM) were used for the classification. The evaluation and the analysis proved that rule-based classifier well performed with 80.36% accuracy as compared to SVM with accuracy 74.0% and NBC with accuracy 71%.

The aforementioned research as summarized in Table 2.1, either used patterns, machine learning models, POS tags, or UMLS (Semantic) mapping for recommendation statement identification. Each approach has its pros and cons. For example, the existing pattern-based identification used single patterns (heuristic, POS, or UMLS) which depends on the extracted patterns and faces difficulties while generalizing the patterns. Also, the pattern extraction required tremendous human time and effort. However, to mitigate these limitations and to get generalized patterns, we need a mixed-method approach, which combines multiple techniques. Therefore, we proposed a machine learning-assisted pattern-based approach by combining heuristic patterns, POS patterns, and UMLS patterns. The mixed-method approach increases the chance of accurate detection of recommendation sentences by more complete and synergistic utilization of various patterns.

Table 2.1: Summary of text classification approaches

Reference	Approach	Type	Learning Method	Algorithms
[45]	Pattern based	Semantic patterns	Manual	-
[46]	Pattern based	Linguistic and Semantic patterns	Manual	-
[47]	ML based	POS tags	Automatic	ZeroR, Naive Bayes, J48, Random Forest
[48]	ML based	tokens with POS tags	Automatic	Rule based, Naive Bayes, SVM

2.2.2 Causality Mining Related Work

Causality mining as an application of causality detection is typically based on two tasks, which includes identification of causal triggers, and causal pairs participating in each relationship [35]. Also known as causal connectives; causal triggers are transitive verbs which form a bridge between causality concepts and identify the cause and its effect. Leveraging the sentence structuring in English language [49], typical causality relation identification methodologies, found in research literature, follow the Noun Phrase (NP) - Verb (V) - NP pattern which corresponds to either Cause - Trigger - Effect or Effect - Trigger - Cause forms ($\langle S \rightarrow NP\text{-Cause}, V \rightarrow \text{Verb-Trigger}, O \rightarrow NP\text{-Effect} \rangle$) [36]. Based on this heuristic, Kaplan and Berry-Bogge [50] provided an early model for creating and using handcrafted linguistic template for causality detection. Kalpana Raja et al. [51], built upon the same idea in addition to identifying and organizing a dictionary based on causal trigger keywords, which was then used to define patterns for causality detection. R. Girju et al. [52] refined the process of identifying the causal verbs by utilizing the WordNet dictionary [53]. Bui et al. [54] applied rule based approach for causal relation extraction on HIV drug resistance. Cole et al. [55] utilized a syntactic parser to convert the SVO structures into SVO triples, which were then passed through various rule based filters for causality detection. S. Zhao et al. [35], pointed towards the existence of diversity in the manner each causal trigger expresses causality. However, the syntactic structure of causal sentences and the way the trigger invokes the causality, can provide satisfactory categorization of the causal triggers, enabling smart application of the causality identification filters. Son Doan et al. [56] presented an application of causal mining by marking several verbs and nouns as causal triggers for extracting causal relations from twitter messages. Saud Alashri et al. [57] proposed a snowball strategy, where the authors defined few causal verbs as "seeds" and enlarged the seed list from climate new text by generalizing the seed verbs. Girju and Moldovan [36] proposed a semi-supervised approach towards causality relation

identification by using the underlying linguistic patterns of the corpus.

Many other automatic causal pattern identification methodologies have relied on the evolution of machine learning models [58]. In particular, [59] has presented a causal relation extraction model using unsupervised learning to detect the noun phrases corresponding to the subject and object of the sentence. By analysing an unannotated raw corpus and using Expected Maximization (EM) along with a Naive Bayes classifier, the authors were able to precisely identify 81.29% of causal relations.

On the other hand, E. Blanco et al. [60] utilized a supervised machine learning approach by first annotating ternary instances as being a causal relation or not, and then applied Bagging with C4.5 decision trees to achieve a precision of over 95% in causal relations and above 86% in non causal ones. These and many other machine learning approaches have been comprehensively classified by [61], which indicates a general trend towards the utilizing of the same, as the models become more mature and stable. Of particular interest are the word embedding methods, which due to their requirement of unsupervised data, scalability, and accuracy have piqued the interest of the NLP research community.

Several initiatives have already led to the state-of-the-art results in completing NLP tasks such as sentiment analysis, text classification, topic modeling, and relation extraction [62]. Zeng et al. [63] classified relations in the SemEval Task 8 dataset using deep convolution neural networks (CNNs). Nguyen et al. [64] introduced positional embedding to the input sentence vector in CNNs for improved relation extraction. Silva et al. [62] proposed a deep learning (CNN) based causality extraction methodology that can detect causality along with its direction. The author addressed the causality detection problem as a three class classification problem, where class 1 indicates the annotated pairs has causal relation with direction entity1 \rightarrow entity2, class 2 implies the causal relation has the direction entity2 \rightarrow entity1, and class 3 entities are non-causal.

Ning An et al. [65] has utilized a word embedding with cosine similarity based approach, which uses an initial causal seed list to identify the causal relationships as a multi-class (four-class) classification problem. With one-hot encoding the authors, convert the causal verbs in the seed list and the verbs identified in Noun Phrase(NP)-Verb Phrase(VP)-Noun Phrase(NP) ternary(triples) into encoding vectors. These vectors are then converted into Embedding vectors using Continuous

Skip-Gram based on a Wikipedia dataset of 3.7 million articles. Finally the encoded vectors are then compared using cosine similarity and the pair with maximum similarity above a pre-defined threshold value of 0.5 are used to classify the causal relationship and evolve the seed list. This method achieved an average F-score of 78.67%. While this methodology presents a significant improvement on previous research initiatives towards causal relationship identification, it suffers from low accuracy, due to its focus on causal verb identification based on a small initial seed list and its limited extension, and classification based, solely on these verbs meanwhile losing context of the causal phrase.

Table 2.2: Summary of causality mining approaches

Reference	Approach	Considered Term	Method
[51]	Dictionary and Rule based	Triggers	Manual
[52]	Lexico-syntactic pattern	Triggers	Manual
[54]	Dictionary and Rule based	Triples	Manual
[55]	Rule based	Triples	Manual
[56]	Pattern based	Syntactic relations	Manual from dependency parser
[57]	Seed words	concept generalization technique	Causal Chains Construction
[35]	ML based	Common features & causal connective features	Naive Bayes
[59]	ML based	Expected Maximization	Naive Bayes
[60]	ML based	Triples - Bagging	C4.5 decision trees
[62]	Deep Learning	Triples	CNN
[65]	word Embedding	Triggers	Skip gram

The summary of literature for causality mining is shown in Table 2.2. Most of existing work used pattern/rule based approach for the causal-effect relation identification task. Also, most of the studies only considered triggers which are manual identified or auto extracted for the classification. The static patterns/rules and seed list may not effectively extract the target relation and demands a dynamic solution such as the proposed methodology. Where the causality mining is performed using the combination of transfer and active learning methodologies.

2.2.3 Clinical Concepts' Value Extraction

Clinical concepts' values such as body weight provide essential information for making clinical diagnosis and treatment decisions. However, manual screening text such as clinical notes for these values is time-consuming and error-prone. Various researchers have proposed automatic techniques for extracting these values. Murtaugh et al. [25] proposed a regular extraction based body-

weight values extraction algorithm known as Regular Expression Discovery Extractor (REDEx). The algorithm extracted patterns from the training dataset by splitting each training sentence into before label segment (BLS), labeled segment (LB), and after label segment (ALS). The BLS and ALS are processed for generalizing punctuation, digits, and white spaces. The lengthy phrase between BLS and ALS are repeatedly trimmed to make their length equal. The BLS and ALS are trimmed from beginning and end, respectively, until a false-positive instance occurs. Finally, the resultant BLS, LS, and ALS phrases are combined as a pattern. The extracted patterns are then used to detect body weight related measures such as weight, height, BMI, etc. in unseen clinical text. The primary limitation of the REDEx is the lack of generalization. The algorithm uses exact concepts/terms in the patterns, while in real textual data a term can be represented by other alternatives.

Redd et al. [66] proposed an update in the REDEx algorithm to overcome its limitation of exact concepts/term usage in the patterns. The authors enlarged the coverage of the REDEx algorithm by generalizing the extracted patterns for Metabolic Equivalents (METs) value extraction. They improved the REDEx algorithm, where concepts of each pattern were replaced with its equivalent length regular expression. The algorithm marked some improvements over REDEx, however, concepts can have variant length alternatives that can not be captured.

Zheng et al. [67] formulate the value extraction as a sequence tagging task. The authors used a joint model of BiLSTM and conditional random fields and named it OpenTag. The LSTM model detects the context and semantics of the text while CRF enforces tagging consistency. They also proposed a novel attention mechanism to get the decision logic of the black box LSTM model. The model provides a detailed explanation for the decision using the proposed attention. Also, the study employed active learning to enhance the model performance over time.

Cai et al. [68] proposed a simple and an effective tool EXTEND for extracting numerical clinical data from narrative notes. The EXTEND processes clinical notes in three simple steps. First, the clinical notes are pre-processed by performing normalization and tokenization of the sentences. Second, the concepts of interest are identified in the input clinical notes. A dictionary of concepts is manually curated that is used to find concepts of interest in the text. Finally, the data extraction is performed by evaluating concepts neighbor tokens for numerical values. The ex-

tracted values are validated via rule base approach. The tool achieved a sensitivity and specificity of 0.95 to 1.0 and an F1 score of 0.92 to 0.96. The major limitation of the EXTEND tool is manual work required for building concepts dictionary which is used to locate concepts of interest in the input clinical case.



Healthcare organizations have seen explosive health data growth and continue producing loads of data, mostly in unstructured textual format. The generated data withholds valuable insights and knowledge. However, machines cannot process and understand the precious data due to its unstructured format, which necessitates an AI-based methodology to translate it to machine-understandable format. This dissertation proposed an end-to-end methodology for unstructured clinical data processing. The workflow of the proposed methodology is shown in Figure 3.1. The methodology processes clinical text in three steps/solutions where each step utilizes the output of its predecessor and enhances the process for its successor steps. Solution one prepares the input text into subsequent components ready format by applying text pre-processing steps. Also, it evaluates each pre-processed sentence of the text for importance, based on the content of the sentence, and classifies it as recommendation or non-recommendation sentence. The non-recommendation sentences are filtered out, and recommendation sentences are passed to solution two for processing.

Solution two takes recommendation sentences as input and produces causal-effect triples as output. The recommendation sentences are processed to get triples of the form $\langle \text{NP}, \text{VP}, \text{NP} \rangle$ where first NP represents noun phrase or subject, second NP represents object phrase, and VP represents verb phrase encapsulated between noun phrases of the sentence. The acquired triples are converted into triple phrases by concatenating the elements of each triple by a space character. The triple phrases are transformed into vectors by using state-of-the-art pre-trained BERT models. The embedded test triples are compared with embedded causal triples for similarity. Test triples having similarities greater than a threshold are tagged as causal triples, otherwise non-causal. The causal classified triples are used by solution three for rule generation as well as for the training

triples evolution utilizing active learning methodology. A domain expert evaluates the causal tagged triples to validate their correctness. The correct identified casual triples are feedback to the trained model to learn it for subsequent runs. The process enhances and increases the model performance over different runs.

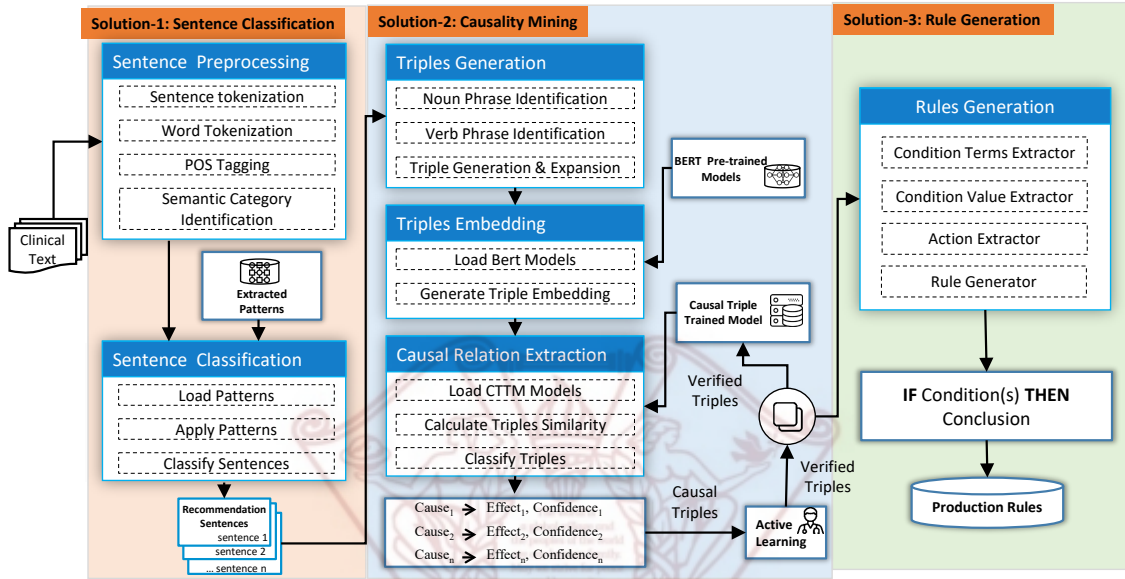


Figure 3.1: The proposed methodology workflow

Solution three aims to transform the output of solution two (cause-effect triples) into executable and machine-understandable knowledge (production rules). As production rules represent knowledge in IF condition(s) THEN conclusion where condition follows the “Key Operator Value”, therefore, we map the identified cause into the key thus requiring operator and value. The proposed solution three identifies the operator and value for the identified entity (key) using its neighbor tokens. We apply the combination of the Stanford NLP parser and “Quantitative Concepts” of UMLS in the neighbors of the target entity and attach it as a possible value that completes the condition part of the rule. The same process will continue for all identified casual triples which result in a set of rules as knowledge. The detail of each solution is presented in the following sections.

3.1 Clinical Text Classification

Text classification aims to evaluate the content of the text and categorize it into various categories. The text presenting similar content will be group together. In the clinical domain text content can broadly be categorized into two groups, one presenting background information while the second group deals with various symptoms, diseases, treatment procedures, and follow-up related information. Among these two groups the latter group is more important and precious for domain experts. However, it is very time consuming and tedious task to manually process the clinic text for classifying the content into aforementioned two groups. The NLP techniques of the current era are capable to overcome the burden by automatic classification, which effectively evaluate and classify the content according to the information carried by the text.

The NLP techniques for text classification can broadly be grouped into three categories, namely pattern-based approach, traditional machine learning approach, and deep learning approach as shown in Figure 3.2. The pattern-based approach relies on handcrafted or auto extracted rules which are used for text classification. The traditional ML-based approach transforms the input text into a structured format followed by machine learning model training. Where the ML model tries to identify insights and logic that can be used later on to distinguish between recommendation and non-recommendation content in unseen text. Similarly, the deep learning methods are inspired by human neurons and try to mimic human-like reasoning and classification capability for differentiating between important and less-important text.

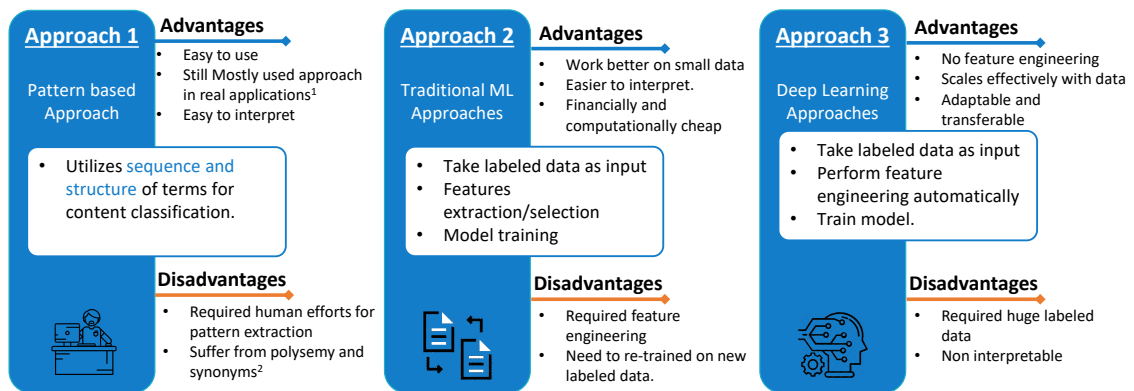


Figure 3.2: Text classification approaches with pros and cons

The pattern-based approach utilizes the sequence and structure of text tokens to build patterns for classification. This approach is widely used in the real field and generally performs better than other approaches. Normally, patterns are extracted by human experts which required a lot of effort and time that ruins the advantages of this approach. Even after a tremendous effort, the extracted patterns face generalization issues. The text intended for classification may use different sequences and terms for expressing the same content. To overcome these issues, this dissertation proposed a machine learning assisted pattern extraction methodology that facilitate human experts for better pattern extraction [69]. The resultant patterns reflect human intelligence as well as artificial intelligence, which can better perform the text classification task. Also, we extend the extracted patterns by substituting its tokens/concept/terms with POS and UMLS tags which overcome the problem of generalization in the patterns. Additionally, an automatic algorithm for semantic pattern extraction is proposed, that can extract patterns from the text without human interventions [70]. The algorithm identifies each token semantic category from UMLS and evaluates the occurrence of various semantic categories. The frequent occurred semantic categories along with its neighbor tokens' semantic are used for semantic patterns generation which can then classify the unseen clinical text effectively.

Machine learning has shown its significance in providing advanced healthcare solutions and revolutionizing the future of filtering a huge amount of textual content [71]. These solutions mitigate the manual effort and improve the text classification performance by learning the classification logic from data, automatically. It required a set of data annotated with their class labels. The training text is first transformed into a structured format, followed by ML model training. The ML model tries to learn the classification logic from the features represented in a structured form. As the model performance can greatly be influenced by the number and quality of features, therefore, this dissertation investigates the effect of features enrichment [72]. We performed feature expansion through synonyms and add semantic features to the feature list. The synonyms were identified from the WordNet dictionary and the semantics of the tokens were added from the UMLS dictionary. The enriched features enable ML Models to better learn the classification task.

Recent years have witnessed tremendous accomplishments via deep learning methods in various fields. The deep learning approach automates the features engineering of traditional ML

approaches. However, these approaches required a huge labeled data compare to traditional machine learning models. While in the clinical domain limited annotated data is available, and the distribution between recommendation and non-recommendation text is also very biased towards non-recommendation. Therefore, this thesis investigates the effect of deep learning methods in the clinical text classification task. As mentioned earlier, the clinical domain has limited training data, therefore, we explore the applications of these advanced models with a large dataset by bootstrapping dataset. The effect of various data augmentation methods on the deep learning model performance is evaluated as presented in Chapter 7.

3.2 Causality Mining

Clinical concepts and their relationships identification is a key cohesive ingredient of any NLP-based clinical solutions to locate clinical terms used in the text along with their effect on other terms. One of the most important relations of interest for clinicians is the cause-effect relationship between concepts. Identification of cause-effect relations is a complex task and required advanced NLP techniques because it is not always written in IF condition THEN conclusion where conditions represent a clinical aspect while the conclusion is the effect of that aspect. A lot of research work has been carried out to accurately perform the causality mining [73]. However, researchers mainly followed a trigger-based approach and focused on causal triggers identification which connects two concepts with cause-effect or effect-cause relationship [51]. The triggers are either manually initialized or learned from training data, later on, the unseen text is evaluated to find any triggers to build the causal relation. This dissertation investigated the phrase embedding-based method for causality mining [74]. The steps performed for causal relation extraction are shown in Figure 3.3. We extract, expend, and embed the extract causal triples from training text. The generated causal embeddings are then used to match with similar clinical concepts withholds causality relationship in the unseen clinical text.

Real world textual data is considered dirty since it contains many defacto linguistic elements which may be a part of daily conversations and routine usage between humans but are not understandable by a computing device. The primary aim of pre-processing is to prepare clinical text for causal phrase extraction which are then used by the subsequent modules. The pre-processing step

cleans the text and normalized it by removing syntactic problems such as redundant text, unrelated information (Explanations, such as this one, in parenthesis which are useful for readers but not required for establishing context), and special characters (-, +, _, etc.) using regular expression. Each processed sentence is then tokenized into words using NLTK word tokenizer. Finally, Part Of Speech (POS) tagging is applied on each word using Stanford CoreNLP Parser [75].

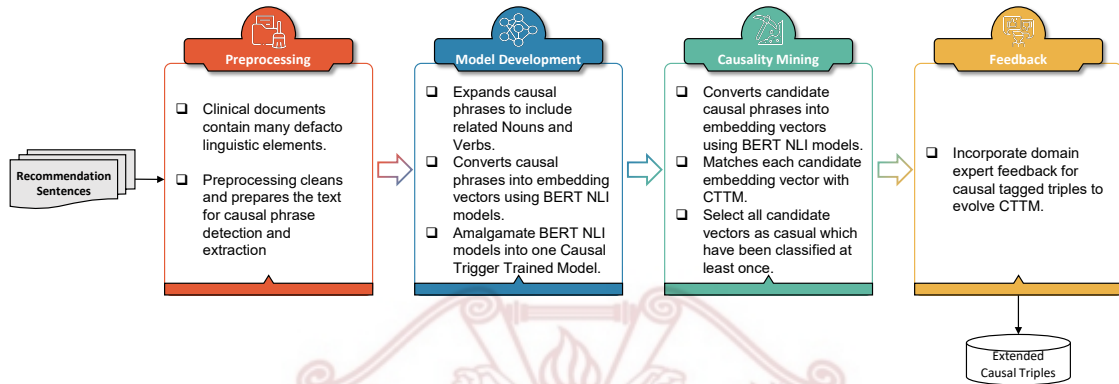


Figure 3.3: Causality mining abstract view

Model Development module generates casual triples of the form <Noun Phrase (NP), Verb Phrase (VP), Noun Phrase(NP)> which can corresponds to either <Cause, Causal Trigger, Effect> or <Effect, Causal Trigger, Cause>. The elements (NP and VP) of the extracted triples are expanded using transfer learning technique on a pre-trained model. This increases the number of causal triples, which in turn increases the scope of causal sentences that can be correctly classified in the testing phase. The acquired set of causal triples are converted into embedding vectors using pre-trained BERT language models. The generated embedding vectors are stored and referred to as Casual Triple Trained Model (CTTM) and used to identify similar causal concepts in the unseen clinical text.

The causality mining module utilizes CTTM on unseen, pre-processed test data, for classifying candidate phrases as causal or non-causal. First, candidate triples are identified from the text. These candidate triples are obtained by collecting all possible phrases of the form <NP, VP, NP> within each pre-processed sentence. For sentences with more than one verb, the noun phrases with longer dependencies are discarded. This is to maintain the context of the nouns with their nearest verb phrase for matching with our causality identification. The extracted candidate triples

are converted into phrases by concatenating each triple element with a white space character. The triple phrases are converted into embedding vectors via BERT models. Each embedding vector is evaluated with causal triples embedding stored in CTTM in terms of max similarity and classified into a causal or non-causal class. We apply human in the loop (active learning) methodology to enhance the CTTM and casualty mining. The causal classified triples are evaluated by a human expert. In case the human expert agrees with the causal tag of a triple, the triple embedding is feedback and added to CTTM which increases the model performance for upcoming evaluation.

3.3 Rules Generation

The rules generation modules take causal triples, the output of causality mining module, and transform into production rules in IF condition(s) THEN conclusion format. The condition part of each production rule consists of key, operation, value. The causal triples consist of a cause entity which can be mapped into condition key while the effect can be mapped to the conclusion part. However, to complete the rule, we required operation and value for the causal concepts. Therefore, we evaluate the source sentence of the triples for the operator and valued identification. The proposed methodology is inspired by EXTEND [68] and used similar steps for operator, and value extraction. An abstract view of the process required for completing a rule condition is shown in Figure 3.4. We evaluate the neighbor tokens of the causal concept in the source sentence for the relational operator and valued token using the UMLS category of the token and parser information. Also, we evaluate the negation of a concept to adjust the condition operator. Finally, the triple is presented as a production rule.

As described, the main hurdle in transforming causal triples into production rules is the operator and value identification. We used the pre-processed source sentences of the causal triples for completing the required information. Each token of the sentence is already tagged with its POS tag. We extend each token tagging with its corresponding semantic type by looking up the UMLS dictionary. The neighbor tokens having the “Quantitative Concept” category or having relational operator marked by UMLS or Stanford NLP parser [75] are considered as candidate value and operator, respectively. The concept is also evaluated for its negation in the source sentence. The operator is updated in case negation is confirmed. Thus the condition is set to the causal concept

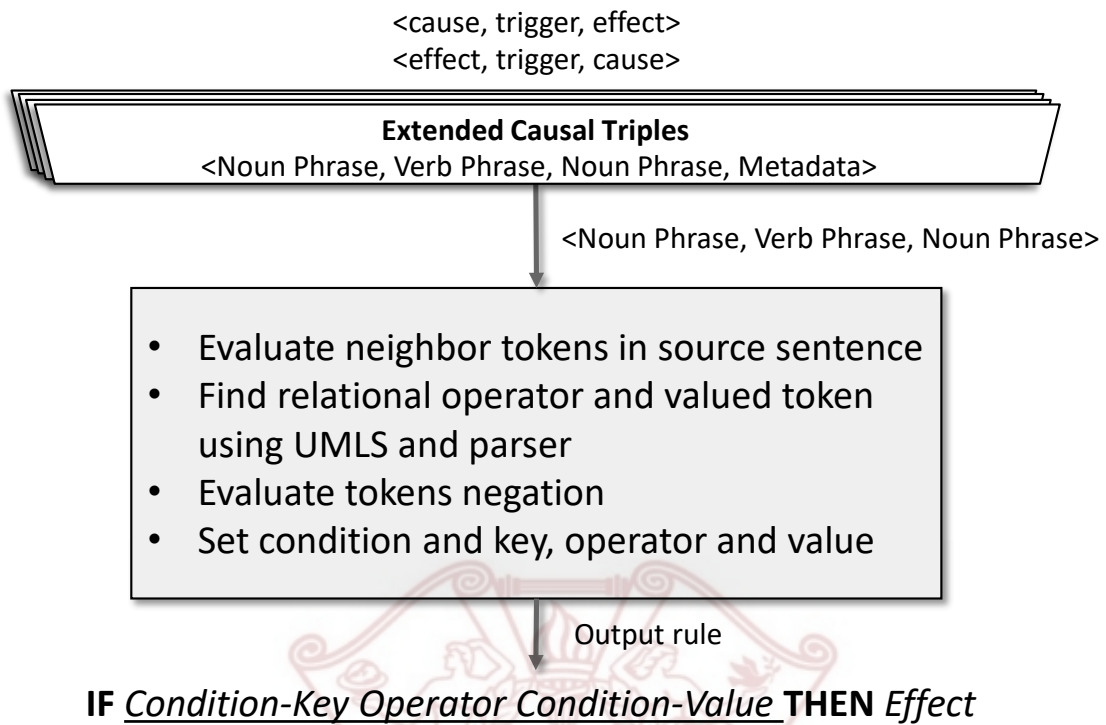


Figure 3.4: Rules generation abstract view

from the triple, operator and value is set as identified which completes the condition part of the rule. The effect concept of the triple is set to the action part of the rule which completes the rule. Rules generated from causal triples with similar source sentence are combined to a single rule. After following the same process for all triples, the acquired list of rules is evaluated for duplicate and conflicting rules. The duplicates are removed while the conflicts are resolved via human intervention. Thus the resultant rules can be used for assisting clinical decisions by automated systems as well as by human experts to increase their knowledge and intuitions.

Clinical content can be categorized into two parts based on the importance of the provided text. First, the background information, which includes abstract information related to the background and point of view of the authors. Second, the disease-specific information, which elaborates causes, consequences, and actions related to a disease. For instance the sentence, “Hypertension remains one of the most important preventable contributors to disease and death.” represents background information, while, “In the black hypertensive population, including those with diabetes, a calcium channel blocker or thiazide-type diuretic is recommended as initial therapy.” represents disease-specific information, also known as a recommendation sentence. Therefore, the understandability and classification of clinical contents is an important step, before its transformation to computer interpretable format. The primary objective of the text classification is to distinguish between recommendation sentences (RS) and non-recommendation sentences (NRS). The research related to text classification and information extraction has greatly benefited from advancement in computing technologies, producing a plethora of algorithms, tools, and applications, based on pattern-based and machine learning approaches [26–30, 76]. This research explores the effectiveness of pattern based, traditional machine learning and advanced machine learning such as deep learning methodologies for clinical text classification.

4.1 Pattern based Classification

The machine learning techniques of the current era advanced various applications including text classification and information extraction. However, in the clinical domain, most of the NLP tasks including, text processing and information extraction, are still using pattern-based approaches [32]. Generally, pattern-based approaches perform better than machine learning models in clinical

text classification [33]. The patterns are usually extracted manually by human experts based on their heuristics [77], semi-automatic or automatically [77]. However, this dissertation focuses on machine learning assisted pattern extraction methodology as shown in 4.1.

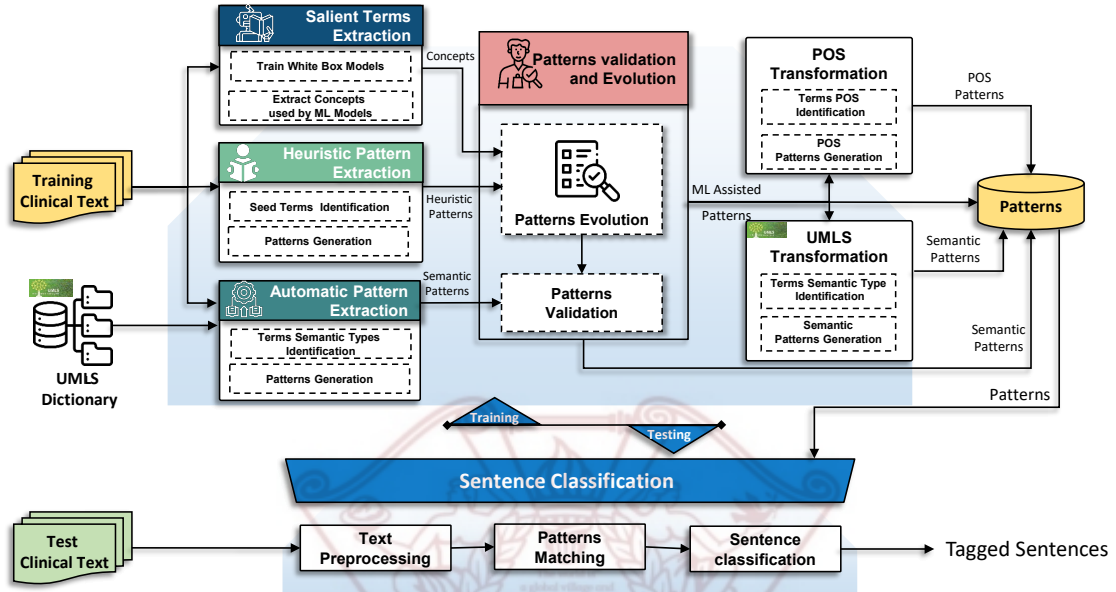


Figure 4.1: The proposed pattern extraction methodology for clinical text classification.

4.1.1 Heuristic Pattern Extraction

The heuristic pattern extraction uses experts heuristics for pattern extraction. The experts focuses on the sequence of terms used in content for patterns generation. We applied a collective decision making effort known as Nominal Group Technique (NGT) to reflect multiple experts heuristics for pattern extraction. Five human experts (Knowledge Engineers KEs) participate in the NGT process. The KEs have more than five years of experience in biomedical text processing, analysis, and pattern extraction. We provided them an annotated clinical text (hypertension guideline [78]) to KEs for extracting patterns based on their heuristics. Heuristic-based decisions are premised on the cognitive ability, rule of thumb, intuitive judgment, an educated guess, and common sense of a person. The following seven steps were performed in the NGT process for extracting the patterns.

1. Introduce all team members and nominate a leader to cordially handle meetings. The annotated CPG is provided to each member, the leader explained the purpose and process of the

study and the voting process.

2. All panel members analyze the provided CPG independently and extract the patterns based on their heuristics that can identify recommendation statements in a CPG.
3. The leader collects all patterns extracted by each member and removes the duplicate patterns. A total of 21 unique patterns were identified by all KEs as shown in Table 4.1.
4. The panel members discuss each pattern and the concerned member explains the reason for selecting the corresponding pattern.
5. All five participants rank each pattern from one to five, where one is the lowest and five being the highest rank. The leader aggregate the ranks of each pattern.
6. A threshold value (total rank ≥ 15) is selected with the consensus of all team members, which is the 60% of team members agreement on a pattern.
7. Select those patterns, which have a higher accumulative rank than the threshold value (15). Based on this criterion, 10 patterns are selected as final patterns shown in Table 4.2.

The key advantage of this approach is, its ease of use and comprehensibility for human beings without detailed domain knowledge. However, this approach required extensive human efforts and time. The resultant patterns vary based on the heuristics of human being involved in the pattern extraction process. To assist experts with machine learning methods we devise machine learning assistive pattern extraction methodology described in the upcoming section.

4.1.2 Machine Learning Assisted Patter Extraction

To assist the manual pattern extraction explained in Section 4.1.1, we applying machine learning techniques for important and salient terms extraction. The concepts and terms utilized by machine learning models for distinguishing recommendation sentences from non-recommendation sentences are considered as salient terms. The objective of machine learning assistance is to identify the key terms in the clinical text, using both supervised and unsupervised machine learning techniques. To extract the salient terms, we transform the clinical text to machine-processable

Table 4.1: Evaluation matrix for nominal group technique (NGT).

S.No	Extracted Patterns	KE-1	KE-2	KE-3	KE-4	KE-5	Total Score
1	.*lead(s)? to.*	3	3	2	1	2	11
2	.*treatment (should with to).*	2	3	4	4	3	16
3	.*initiat(.*).treatment.*	2	3	2	3	4	14
4	.*to improve.*	4	4	1	3	5	17
5	.*evidance(.*)(to)? support.*	1	4	2	1	3	11
6	.*(patient(s)?)? with (disease).*	3	3	4	5	4	19
7	.*should (include continue).*	5	3	3	2	5	18
8	.*appli(es ed)ed (to)?.*	2	1	3	2	3	11
9	.*can be used.*	3	2	2	1	2	10
10	.*(add remove)(.*) drug.*	4	4	3	5	5	21
11	.*(panel)(.*)(recommend(ed)? conclude(ed)? include(d)?).*	4	2	2	1	3	12
12	.*less effective.*	1	3	4	2	3	13
13	.*treatment (does not)? need.*	2	3	2	3	3	13
14	.*regardless of.*	3	4	3	3	2	15
15	.*meet.*goal.*	2	1	2	2	3	10
16	.*(increase decrease).*dose.*	5	4	5	5	4	23
17	.*(recommend(ed)?) treatment.*	3	3	4	3	3	16
18	.*(improve(ment)? high quality).*dose.*	2	2	3	3	2	12
19	.*(Recommendation /d+/s+.*	5	5	5	5	5	25
20	.*expert(s)?.*opinion.*	3	3	2	3	2	13
21	.*(dis)?continu(e ed ing ation).*	4	3	3	2	4	16

Table 4.2: Extracted heuristics patterns

S.No	Patterns without Salient Terms
1	.*(add remove) (.*). drug.*
2	.*(recommend(ed)?) treatment.*
3	.*to improve.*
4	.*(increase decrease) .*dose.*
5	.*treatment (should with to).*
6	.*Recommendation /d+/s+.*
7	.*should (include continue).*
8	.*(dis)?continu(e ed ing ation).*
9	.*regardless of.*
10	.*(patient(s)?)?with (disease).*

format by tokenization, stemming, case transformation, stop word removal, and synonym identification. We trained a set of supervised machine learning models comprising of decision tree

and rule induction, and unsupervised algorithms LDA, and word2vec to find the key contributing terms in a text for taking sentence classification decision. These techniques were selected due to their results transparency and effectiveness in the classification task. We applied various parameter settings for each model to check its classification accuracy and extract the final terms, which are then used for making the classification decision.

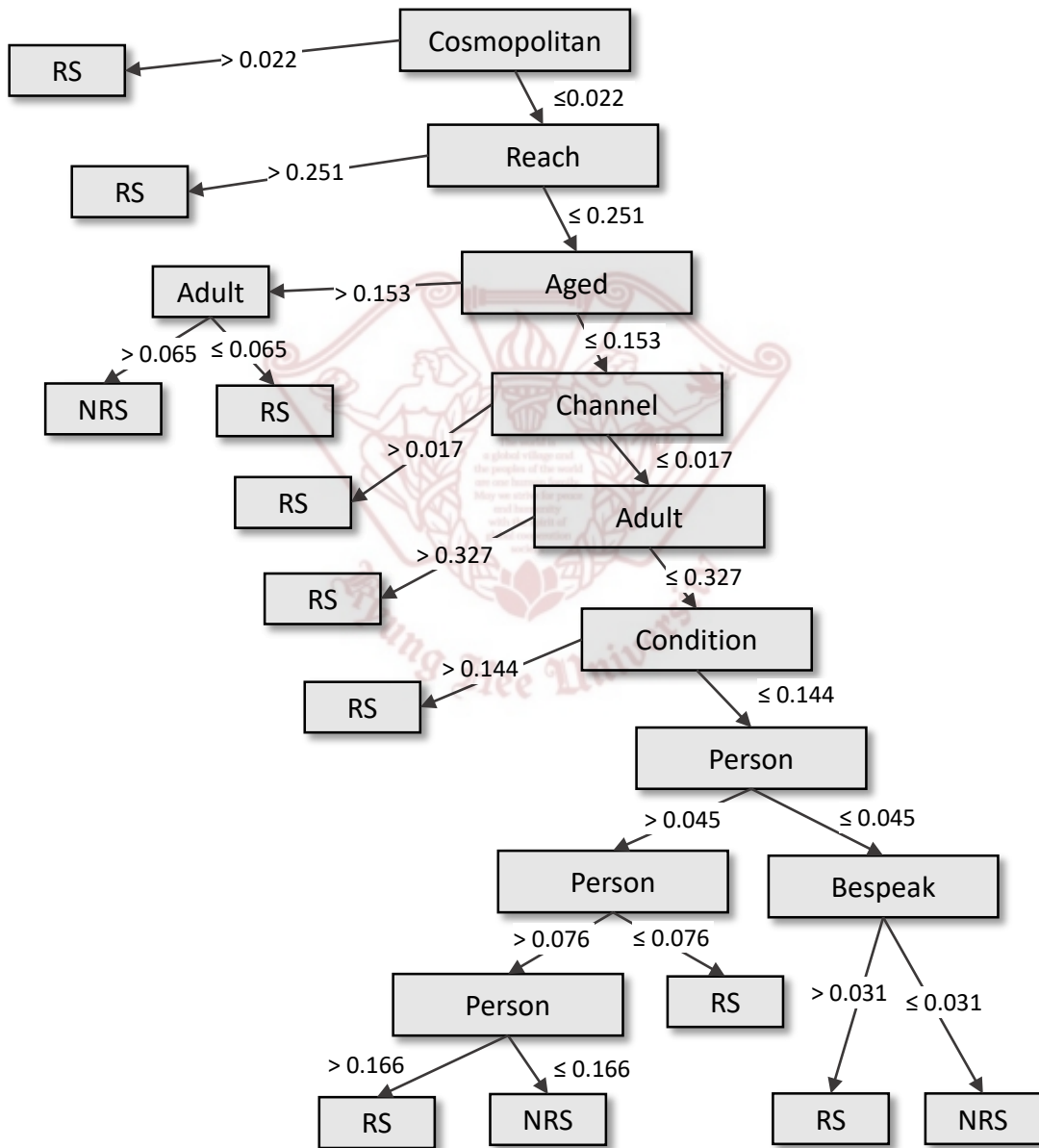


Figure 4.2: Example decision tree model for salient term extraction.

As an example, in the decision tree model, we apply gain ratio, information gain, accuracy, and Gini index splitting criteria. We also evaluate the models' behaviors with and without feature selection. In feature selection, filter-based and wrapper-based techniques were applied to limit the number of features and nodes of the final model by eliminating irrelevant features. While, identifying the correct number of features is still an open research issue, in this study, we used the grid search technique [79] to dynamically set the number of features for a model. We check the terms considered by the model, generated after feature selection to get a list of salient terms considered by the model.

Table 4.3: List of salient terms considered by machine learning models

S.No	Decision Tree	Rule Induction	LDA	Word2vec
1	cosmopolitan	cosmopolitan	goal	recommend
2	angiotensin	reach	low	facilitate
3	bespeak	black	population	improve
4	adult	better	treatment	consideration
5	aged	opinion	year	evidence
6	animation	aged	recommendation	assess
7	condition	condition	evidence	condition
8	reach	former	pharmacological	quality
9	black	case	initiate	regardless
10	decrepit	commend	hypertension	referral

The example of the decision tree model is shown in Figure 4.2. The decision tree model have considered total eight unique salient terms, "cosmopolitan", "reach", "aged", "adult", "channel", "condition", "person", and "bespeak" for distinguishing recommendation sentences from non-recommendation sentences in a text. We considered all terms as salient terms, which are extracted by given models with all possible settings. A list of partial salient terms considered by various machine learning models is given in Table 4.3. Therefore, we apply the NGT process again to evolve the experts extracted heuristic patterns. In this phase of the NGT, we provided a list of salient terms to all KEs and asked them to reevaluate their extracted patterns. The aforementioned steps of NGT were performed again to reexamine the patterns with consideration of salient terms. The KEs modified the extracted patterns based on the salient terms and the final agreed-upon heuristics patterns list is given in Table 4.4. The patterns became more general compared to

patterns without considering salient terms. Most of the selected patterns included some of the salient terms to boarder its scope. As an example the pattern `".*(recommend(ed)?) treatment.*"` became `".*(recommend(ed)? |better) treatment.*"` after reflecting salient term "better" in the pattern.

The key advantage of this approach is, its ease of use and comprehensibility for human beings without detailed domain knowledge. However, this approach highly depends on the terms and terminologies of a specific text. Therefore, the extracted patterns may not well-perform for all text such as guidelines. To overcome this drawback, we generalized the extracted patterns with the incorporation of two other techniques POS and UMLS patterns for getting a generic solution.

Table 4.4: Extracted heuristics patterns with salient terms

S.No	Patterns with Salient Terms
1	<code>".*(give add remove) (.* drug.*"</code>
2	<code>".*([I i]n) (black general) (.* population.*"</code>
3	<code>".*(recommend(ed)? better) treatment.*"</code>
4	<code>".*(increase decrease) .*dose.*"</code>
5	<code>".* ((public)? opinion) .* treatment (should with to).*</code>
6	<code>".*Recommendation /d+/s+:.*</code>
7	<code>".*should (include continue).*</code>
8	<code>".*(dis)?continu(e ed ing ation) reach .* goal .*</code>
9	<code>".* (regardless of) (having age).*</code>
10	<code>".*(patient(s)? adult (population group))?with (disease).*</code>

The general purpose of POS tagger is to briefly characterize and disambiguate the grammatical category of words in a specific context. It helps to find the similarity and distinction between words. In the proposed method, the POS-based classification is used to generalize the solution for avoiding domain dependency. In this study, the application of the POS tag produced inferior results. Therefore, we used the semi-POS method, which is the combination of POS tags along with clue words. For example, in `".* VB .* drug .*" "VB"` is a POS that represents a verb while "drug" is a clue word. The list of POS tags, used in the study, are shown in Table 4.5.

The extracted heuristic patterns shown in Table 4.4 is transformed into POS patterns as shown in Table 4.6. We employed the Stanford CoreNLP parser [75] to parse the input sentences to their POS categories. The input sentences were assessed by matching with the POS tags listed in Table

Table 4.5: List of used POS tags

Tag	Description	Tag	Description
CD	Cardinal number	IN	Preposition/sub-conj
MD	modal	NN	Noun, sign. or mass
JJ	Adjective	TO	'to'
JJR	Adjective, comparative	VBG	Verb present participle
VB	Verb base form	-	-

4.5. The sentences matched with one or more patterns were tagged as RS and NRS, otherwise. Finally, all NRS sentences were filtered out, and RS sentences were left for further processing. The POS-based filter reduced domain dependency and increased the accuracy of our proposed system. Here, the most significant POS tags used for the identification are “Nouns” and “Verbs”.

Table 4.6: List of extracted POS patterns

S.No	Patterns
1	. * VB . * drug . *
2	. * IN . * JJ . * population . *
3	. * (VB JJR) . * treatment . *
4	. * NN . * dose . *
5	. * (JJ)? NN . * treatment (MD IN TO) . *
6	NN(/s+)? : (/s+)? CD . *
7	. * VB (include continue) . *
8	. * (VB+) . * goal . *
9	. * (regardless of) VBG age . *
10	. * (JJ NN) IN disease . *

The heuristic patterns displayed in Table 4.4, are also transformed into UMLS based patterns to achieve further generalization. The UMLS based patterns, also known as semantic patterns, cover a wide range of recommendation sentences. This process, additionally improves the accuracy of the system by identifying the semantics of words and phrases in a sentence to clarify its contextual meaning.

The UMLS is a knowledge source, which contains medical vocabularies, maintained by the US National Library of Medicine [80]. It provides an interface for retrieving biomedical concepts and semantic relations, by integrating a plethora of services, and assisting in biomedical information processing and retrieval. Recommendation sentences mostly contains the biomedical phrases,

which can help to distinguish it from non-recommendation sentences. Using this heuristic, first, we identify the UMLS phrases using a tool called MetaMap [81] which can identify the UMLS concepts behind medical text. Using this information, we map phrases of each sentence with its corresponding biomedical concept. We then extract UMLS patterns by analyzing the tagged sentences, identifiers, and their sequence. The example for one of the extracted patterns is shown in Figure 4.3. A list of UMLS patterns used in our study is shown in Table 4.7. The matched sentences with one or more of the UMLS patterns are finally tagged as RS, and NRS otherwise. The NRS tagged sentences are then filtered out, and RS sentences are stored for further processing.

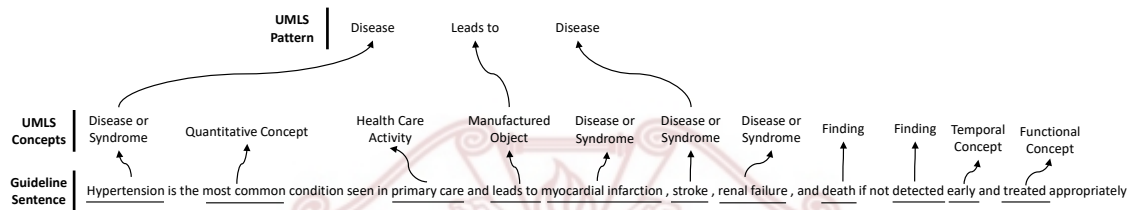


Figure 4.3: Example of UMLS based pattern extraction

Table 4.7: List of extracted UMLS Patterns

S.No	Patterns
1	.*(give add remove) .* Pharmacologic Substance .*
2	.*([I i]n) .* Population Group .*
3	.*(Health Care Activity Qualitative Concept). * Functional Concept.*
4	.*Functional Concept .* Pharmacologic Substance.*
5	.*Qualitative Concept .*Functional Concept (should with to).*
6	.*Idea or Concept /d+/s+.*
7	.* should .* (Functional Concept Idea or Concept).*
8	.*Idea or Concept .* Intellectual Product .*
9	.* regardless of Organism Attribute.*
10	.* Population Group .* with .* (Disease or Syndrome) .*

The extracted patterns (Heuristic, POS, UMLS) shown in Table 4.4, Table 4.6, and Table 4.7, respectively, are used to classify a clinical sentence as RS or NRS. We combine the sentences labeled as RS by heuristic patterns, POS patterns, and UMLS patterns, removing duplicates and finalized the RS tagged sentences and non-recommendation sentences which will be used in latter stages. As mentioned earlier, pattern based approaches performs better, however required exten-

sive human efforts and time which reduces its advantages and limits its utilization. To overcome this limitation, this research proposed an automatic pattern extraction methodology explained in the upcoming section.

4.1.3 Automatic Pattern Extraction

To overcome the manual effort and lack of generalization during manual pattern extraction, we devise a novel algorithm as shown in Algorithm 1. The algorithm identifies and extracts patterns for recognizing RS sentences in clinical text, automatically. Initially, it identifies each token's semantic category/concept by utilizing UMLS dictionary. We represent each token in ('*token*', '*UMLSConcept*') format, where *token* represents textual token while *UMLSConcept* represents token's Semantic type of UMLS. We count the occurrences of each concept in the text and find the list of initial candidate concepts $C = [c_1, c_2, \dots, c_n]$. A concept c_i where $i = 1, 2, \dots, n$ is considered as candidate concept if it is used more than a defined threshold CT value in a given text i.e when $count(c_i) \geq CT$. The concepts also depends on its context and neighbors concepts. Therefore, we generate a context window for each candidate concept. The context window cw of a candidate concept c_i is $cw_i = [c_{i-n}, \dots, c_{i-2}, c_{i-1}, c_i, c_{i+1}, c_{i+2}, \dots, c_{i+n}]$ for $i = 1, 2, \dots, n$, where $c_{i-n}, \dots, c_{i-2}, c_{i-1}$ represents the preceding concepts while $c_{i+1}, c_{i+2}, \dots, c_{i+n}$ represents the succeeding concepts of a candidate concept c_i . We filter out ineffective context window based on its occurrence by applying context window threshold CWT i.e $count(cw_i) < CWT$ to restrict the number of patterns and filter out ineffective patterns. The remaining context windows are transform to regular expressions as patterns $P = [p_1, p_2, \dots, p_n]$, where p_i for $i = 1, 2, \dots, n$ represents an extracted pattern i . The extracted patterns utilizes UMLS semantic categories, therefore, the resultant patterns are referred as semantic patterns and used to categorize clinical text into RS and NRS.

The detailed example of the proposed pattern extraction methodology is shown in Figure 4.4. In the example, we have used concept threshold CT as two; therefore, all the concepts that appear less than twice is filtered out. The remaining two concepts having occurrences count greater than or equal to two is the initial candidate concepts. The context window size is selected as three in the example; therefore, we considered one preceding and following concepts of each candidate

Algorithm 1: Automatic pattern extraction algorithm

Input : Training Corpus C , $UMLS$, Concept Threshold CT , Context Window Threshold CWT

Result: Patterns $P = \{p_1, p_2, p_3, \dots, p_n\}$

```

1  foreach document  $d \in C$  do
2      Concepts  $C \leftarrow []$ 
3      Sentences  $S \leftarrow sent\_tokenize(d)$ 
4      foreach sentence  $s_i \in S$  do
5           $s_i \leftarrow s_i.lower()$ 
6          Sentence Concepts  $SC \leftarrow word\_tokenize(s_i)$ 
7           $SC \leftarrow [w_j \text{ for word } w_j \in SC \text{ if } !(w_j \in stopwords.words())]$ 
8           $C.append(SC)$ 
9      end
10     Concept Semantics  $CS \leftarrow []$ 
11     foreach concept  $c_i \in C$  do
12         Concept Semantics  $CS \leftarrow token\_semantics(c_i, UMLS)$ 
13     end
14     Uniques Concepts  $UC \leftarrow Counter(C).keys()$ 
15     Candidate Concepts  $CC \leftarrow []$ 
16     foreach concept  $c_i \in UC$  do
17         if  $count(c_i) \in C > CT$  then
18              $CC.append(c_i)$ 
19         end
20     end
21     context Window  $CW \leftarrow []$ 
22     foreach concept  $C_i \in CC$  do
23          $CW.add([c_{i-n}, \dots, c_{i-2}, c_{i-1}, c_i, c_{i+1}, c_{i+2}, \dots, c_{i+n}])$  where
             $i = 1, 2, 3, \dots, n, c_{i-1}, c_{i-2}, \dots, C_{i-n}$  Represents the preceding concents and
             $c_{i+1}, c_{i+2}, \dots, c_{i+n}$  represents the succeeding concept of a candiate concetp  $c_i$ 
24     end
25     foreach  $cw_i \in CW$  do
26         if  $count(cw_i) > CWT$  then
27              $P.append(generatePattern(cw_i))$ 
28         end
29     end
30 end
31 return  $P$ 

```

concept. One is selected as the context window threshold; thus, no any context window has been eliminated. Finally, all the context windows are represented in regular expression form as final patterns.

Input Sentence	In the black hypertensive population, including those with diabetes , a calcium channel blocker or thiazide-type diuretic is recommended as initial therapy .
Preprocessed Sentence	'black', 'hypertensive', 'population', 'including', 'diabetes', 'calcium', 'channel', 'blocker' 'thiazide-type' 'diuretic', 'recommended', 'initial', 'therapy'
Tokens, UMLS Concepts	['black', 'Population Group'], ['hypertensive', 'Finding'], ['population', 'Quantitative Concept'], ['including', 'Functional Concept'], ['diabetes', 'Disease or Syndrome'], ['calcium', 'Biologically Active Substance'], ['channel', 'Spatial Concept'], ['blocker', 'Pharmacologic Substance'], ['thiazide-type', 'Pharmacologic Substance'], ['diuretic', 'Pharmacologic Substance'], ['recommended', 'Idea or Concept'], ['initial', 'Temporal Concept'], ['therapy', 'Functional Concept']
Concepts Count	['Population Group' : 1, 'Finding': 1, 'Quantitative Concept' : 1, 'Functional Concept' : 2, 'Disease or Syndrome' : 1, 'Biologically Active Substance' : 1, 'Pharmacologic Substance' : 3, 'Idea or Concept' : 1, 'Temporal Concept' : 1]
Candidate Concepts	['Functional Concept' : 2, 'Pharmacologic Substance' : 3]
Concepts Context Windows	['Quantitative Concept', ' Functional Concept ', 'Disease or Syndrome'], ['Idea or Concept', 'Temporal Concept', ' Functional Concept '], ['Spatial Concept', ' Pharmacologic Substance ', 'Pharmacologic Substance'], ['Pharmacologic Substance', ' Pharmacologic Substance ', 'Pharmacologic Substance'], ['Pharmacologic Substance', ' Pharmacologic Substance ', 'Idea or Concept']
Concepts Context Windows	['Quantitative Concept', ' Functional Concept ', 'Disease or Syndrome'], ['Idea or Concept', 'Temporal Concept', ' Functional Concept '], ['Spatial Concept', ' Pharmacologic Substance ', 'Pharmacologic Substance'], ['Pharmacologic Substance', ' Pharmacologic Substance ', 'Pharmacologic Substance'], ['Pharmacologic Substance', ' Pharmacologic Substance ', 'Idea or Concept']
Final Patterns	[.*(Quantitative Concept).*(Functional Concept).*(Disease or Syndrome).*], [.*(Idea or Concept).*(Temporal Concept).*(Functional Concept).*], [.*(Spatial Concept).*(Pharmacologic Substance).*(Pharmacologic Substance).*], [.*(Pharmacologic Substance).*(Pharmacologic Substance).*(Pharmacologic Substance).*(Pharmacologic Substance).*], [.*(Pharmacologic Substance).*(Pharmacologic Substance).*(Idea or Concept).*]

Figure 4.4: Semantic pattern extraction example

4.2 Traditional Machine Learning based Classification

The primary focus of this research is to automatically and efficiently extract recommendation statements from a clinical text and filter out background information using machine learning algorithms. To achieve this goal, we devised an NLP pipeline as depicted in Figure 4.5. The proposed pipeline accomplished the aforementioned goal in two major steps, it transforms a CPG into a structured format (Word Vector) and then trained an ensemble learning model that uses the base classifier including Naïve Base, Generalized Liner Model, Random Forest, Deep Learning, and

Decision Tree on the generated structured document.

Initially, the clinical sentences along with label are loaded to working space. We used the Term Frequency-Inverse Document Frequency (TF-IDF) scheme for the creation of word vector. The test is prepared for machine processing via various operation ranging from tokenization to synonyms identification. The input text is split into tokens based on word spacing scheme. The words of each token are then transformed to its base format using WordNet stemming followed by Transform Case which converts all tokens to its lower case to maintain symmetry. Some of the word tokens despite maximum usage in the document may have limited impact known as stop words removed by filter stopwords. We applied the Part-of-Speech (POS) using PENN Tree Scheme, employed pattern (NN—NB) to filter the names and verbs used in the input text. As we notice that the clinical recommendation statements mostly consists of disease/medicine name and action on them. We also employed the word expansion mechanism to make the word vector more comprehensive for effective classification. For word expansions, we added synonyms component to the pipeline. We used WordNet dictionary for synonyms identification [82]. We used MeaningCloud services to find and extract aspects of the input text. We created the local copy of UMLS dictionary at MeaningCloud and then used the APIs services for the aspects/concepts extraction based on the created dictionary. The aspects/concepts addition to the data increased the performance of basic classifier as well as ensemble learning base classification as discussed in the result section. The final outcome of this process is a structured data (word vector) consists of all tokens of interest along with synonyms and their aspects/concepts. This document/structured data will be then used for training the machine learning algorithms.

Ensemble learning combines and applies multiple models on the same instance of data to accurately predict the class label for this instance to reach the final conclusion. The algorithm considered in this study includes Naïve Bayes, Generalized Liner Model, Random Forest, Deep Learning, and Decision Tree. The majority voting technique was used to get the final decision. In this technique, the computed results of each algorithm are analyzed in order to determine the final class recommended by most of the algorithms. The trained model is then used to classify unseen text statements to recommendation statement (RS) or non-recommendation statement (NRS).

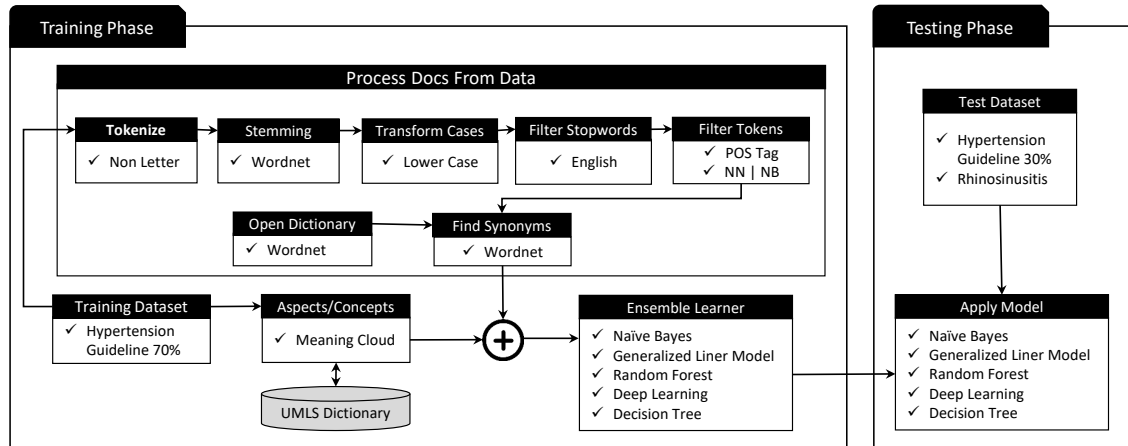


Figure 4.5: Machine learning pipeline for text classification

4.3 Deep Learning based Classification

Deep learning has tremendously improved the application of various domain with no exception to medical domain. However, the primary limitation of these models are, they are data hungry models and required a bunch of annotated data for training. While in medical domain data annotation is very complex and expensive which limit the application of these advance models. Also, the number of classes are very skewed which biased the model towards the majority class. Clinical experts normally interested in minority classes. This research focuses on these two issues of the deep learning model to take the advantages of the available advanced AI models. We explored the application of data augmentation and bootstrapping for the applications of the deep learning models [83]. We expend the available limited annotated data with skewed class distribution with various data augmentation techniques including merging various dataset, swapping token positions and replace tokens with their synonyms.

In the merged data we generate a dataset by combining all existing annotated data set (three in our case) into a single set. The resultant data set is used for training and testing the deep learning models. However, as all the datasets are inclined towards non-recommendation sentences, therefore, the resultant models are also biased toward the non-recommendation sentences. We cover this limitation by duplicating the recommendation sentences and replacing their token position. The resultant data set called swap data is used for the model training and evaluation after split-

ting into training and test slices. Finally, we generated a third dataset by duplicating the existing recommendation sentences and replacing their tokens with synonyms using WordNet dictionary. The resultant dataset called augmented data is evaluated for the application of deep learning models. The evaluation of various large dataset using data augmentation techniques enable us to get benefits of the advance AI models such as deep learning even with limited annotated data.



Modern medicine and healthcare services have greatly improved the daily human life and yet they are beleaguered by constant evolution of diseases, newfound scientific discoveries, and state-of-the-art engineering inventions. This evolution necessitates the use of information technology in general and natural language processing in particular to mine the plethora of healthcare data, information, and knowledge sources to form computable resources. As a part of this endeavor, we present a framework and its novel application for automatically detecting and classifying causal relationships in healthcare textual data. The framework processes clinical text such as clinical notes and clinical practice guidelines, to extract causal knowledge for enabling the medical experts to perform effective diagnosis, treatment, and follow up.

The framework provide four main service categories/modules; Preprocessing, Model Development (MD), Causality Mining (CM), and Feedback Loop as depicted in Figure 5.1. The algorithmic steps of the methodology are shown in Algorithm 2. The preprocessing module transforms the input textual corpora into syntactic enriched sentences which are used by both MD and CM modules for training and applying casual relationship identification model, respectively. The MD module extracts causal triples from the annotated dataset and uses various pre-trained models to self-expand and then generate embedding vectors forming the Causal Trigger Trained Model (CTTM). This model is then used to mine candidate causal relations from unseen clinical text by the CM module, subsequently preparing the causal relationships for verification by an expert. A feedback loop based on the experts' assessment towards the correctness of each relationship, is passed to MD for actively improving the CTTM for future applications. Each of these modules is further discussed in the following subsections.

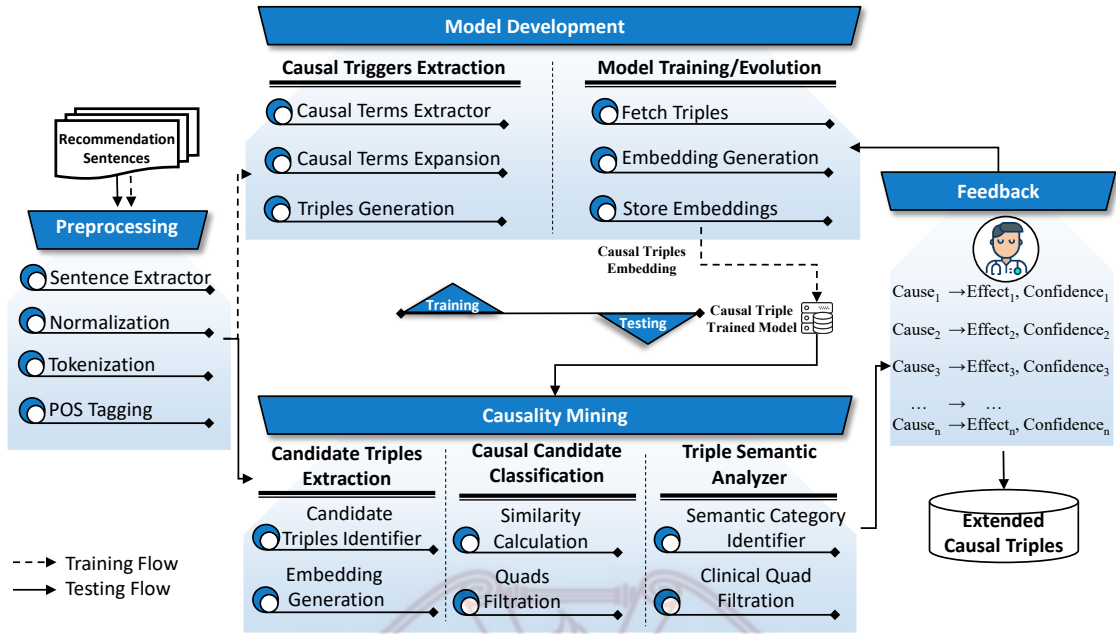


Figure 5.1: Causality mining workflow

5.1 Preprocessing Module

Real world textual data is considered dirty since it contains many defacto linguistic elements which may be a part of daily conversations and routine usage between humans but are not understandable by a computing device. The primary aim of preprocessing is to prepare clinical text for causal phrase extraction which are then used by the MD module to expand the list of causal triggers and by the CM module for semantic comparisons.

The first step of this process is to extract individual sentences from the input corpora using the Natural Language Toolkit (NLTK) [84] sentence tokenizer. Syntactic problems such as redundant text, unrelated information (Explanations, such as this one, in parenthesis which are useful for readers but not required for establishing context), and special characters (-, +, -, etc) are removed in the normalization step using regular expression. Each processed sentence is then tokenized into words using NLTK word tokenizer. Finally, Part Of Speech (POS) tagging is applied on each word using Stanford CoreNLP Parser (version 3.9.2) [75], thereby completing the preprocessing stage. The syntactically enriched sentences are now ready for causal phrase extraction by the MD module and semantic comparisons by the CM modules.

Algorithm 2: Proposed causality mining algorithm

Input : Clinical Documents D , $CTTM$
Result: Causal Medical Quad $MQ = \{q_1, q_2, q_3, \dots, q_n\}$

- 1 Bert Models =
 $M = \{nli-base-mean-tokens, nli-large-mean-tokens, nli-base-max-tokens, nli-large-max-tokens, nli-base-cls-token, nli-large-cls-token\}$
- foreach** document $d_i \in D$ **do**
- 2 Triples $T \leftarrow []$
- 3 Sentences $S \leftarrow sent_tokenize(d_i)$
- 4 **foreach** sentence $s_i \in S$ **do**
- 5 $s_i \leftarrow remove_words_in_brackets(s_i)$
- 6 $s_i \leftarrow replace_abbreviation(s_i)$
- 7 $s_i \leftarrow normalize(s_i)$
- 8 $tokens \leftarrow word_tokenize(s_i)$
- 9 $POS_tokens \leftarrow Pos_tag(s_i)$
- 10 $T.append(generate_triple(pos_tokens))$
- 11 **end**
- 12 Causal Quads $CQ \leftarrow []$
- 13 **foreach** triple t_i in T **do**
- 14 **foreach** model m_i in M **do**
- 15 embedding vector $ev \leftarrow embed(t_i, m_i)$
- 16 $similarity \leftarrow max(similarity(ev, m_i, CTTM))$
- 17 **if** $similarity > m_\alpha$ **then**
- 18 $CQ.append(< NP, VP, NP, similarity >)$
- 19 **end**
- 20 **end**
- 21 **end**
- 22 Medical Quads $MQ \leftarrow []$
- 23 **foreach** $cq_i \in CQ$ **do**
- 24 $concept_1 \leftarrow get_concept(cq_i, 1)$
- 25 $category_1 \leftarrow get_category(concept_1)$
- 26 $concept_2 \leftarrow get_concept(cq_i, 2)$
- 27 $category_2 \leftarrow get_category(concept_2)$
- 28 **if** $category_1 \neq Null$ AND $category_2 \neq Null$ **then**
- 29 $MQ.append(cq_i)$
- 30 **end**
- 31 **end**
- 32 **end**
- 33 **return** MQ

5.2 Model Development (MD) Module

The MD module extracts an initial casual trigger list from the syntactically annotated data produced via preprocessing of the training dataset. This list is then expanded using pre-trained models, before being converted into embedded vectors and becoming a part of the CTTM. This process completes in two steps, Causality Trigger Extraction and Model Training/Evolution, which are discussed in the following sub-sections.

5.2.1 Causal Trigger Extraction

In stage one causal trigger extraction is used to generate a causal triple of the form $\langle NP, VP, NP \rangle$ which can corresponds to either $\langle \text{Cause}, \text{Causal Trigger}, \text{Effect} \rangle$ or $\langle \text{Effect}, \text{Causal Trigger}, \text{Cause} \rangle$. This process starts by extracting causal triggers which appear as a combination of these noun phrases and verbs from syntactically enriched sentences (while there may be other sentence structures corresponding to causal relationships, in this research we are only focused on processing the aforementioned structures). Since there could be many verbs within each noun, and there can be multiple phrases within each sentence that qualify as a causal triple, we collect the set of all verbs within well-defined noun phrases. We then expand the elements (NP and VP) of the causal triple using transfer learning technique on a pre-trained model. In the presented approach, we have applied transfer learning using the pre-trained Google News model, which can be replaced with by utilizing other expansion techniques such as, synonym search from WordNet dictionary [53], ConceptNet Numberbatch Model [85], and/or Facebook Fasttext Model [86].

The expansion of each term is restricted to top ten similar words. This choice of selecting only the top ten similar words is driven by the impact of this selection on quantity of operations required for embedding vector generation and their subsequent comparisons.

Once the triples have been expanded, we then apply Cartesian product between the two expanded noun phrases (Expansion set of the 1st and 3rd element of the causal triple) and one of the verb expansion from the causal triple. This increases the number of causal triples, which in turn increases the scope of causal sentences that can be correctly classified in the testing phase.

5.2.2 Model Training/Evolution

In stage two, the set of causal triples are converted into embedding vectors using pre-trained BERT language models. In order to generate the embedding vectors, the three elements of the causal triple are concatenated by spaces, producing a phrase of the form “*NP V NP*”. The collection of these embedded vectors, forms the Causal Triple Trained Model (CTTM). In our experiments, which will be discussed in later sections, we compared 6 BERT Natural Language Inference(NLI) models with mean, max, and cls tokens [87], in terms of their ability to correctly classify causal sentences, from unseen test dataset. Based on the coverage of causal terms by these models, a multi-model approach is well suited for the causality mining task. As a result, each causal phrase is converted into 6 embedding vectors generated via the 6 BERT NLI models. While the space of the CTTM is increased 6-fold, due to this enhancement, it also provides better semantic matching performance, which will be discussed in the results section.

5.2.3 Example of Model Development (Training Phase)

An example of this process is shown in Figure 5.2. Starting with a sample sentence from our training dataset, which contains the annotated cause and effect entities enclosed within e1 and e2 tags in step 1, we applied preprocessing on it. This produced a POS annotated sentence in step 2, which is used to identify the tagged nouns and verb terms between them in step 3. Each noun term is further expanded to include the preceding adjectives, if any. Any verb terms outside the tagged nouns are ignored. As shown in the step 4 “is” and “triggered” are two of the candidate verbs identified in this process, while “disease” and “ingestion” are their encapsulating noun phrases. In step 5, each of the participating noun and verb phrase is expanded by identifying their closely related alternatives. In step 6, we applied Cartesian product on the sets of two nouns and each verb phrase, producing the set of expanded causal triples. In step 7, the causal triples are converted into causal phrases, which are then converted into embedding vectors as shown in step 8.

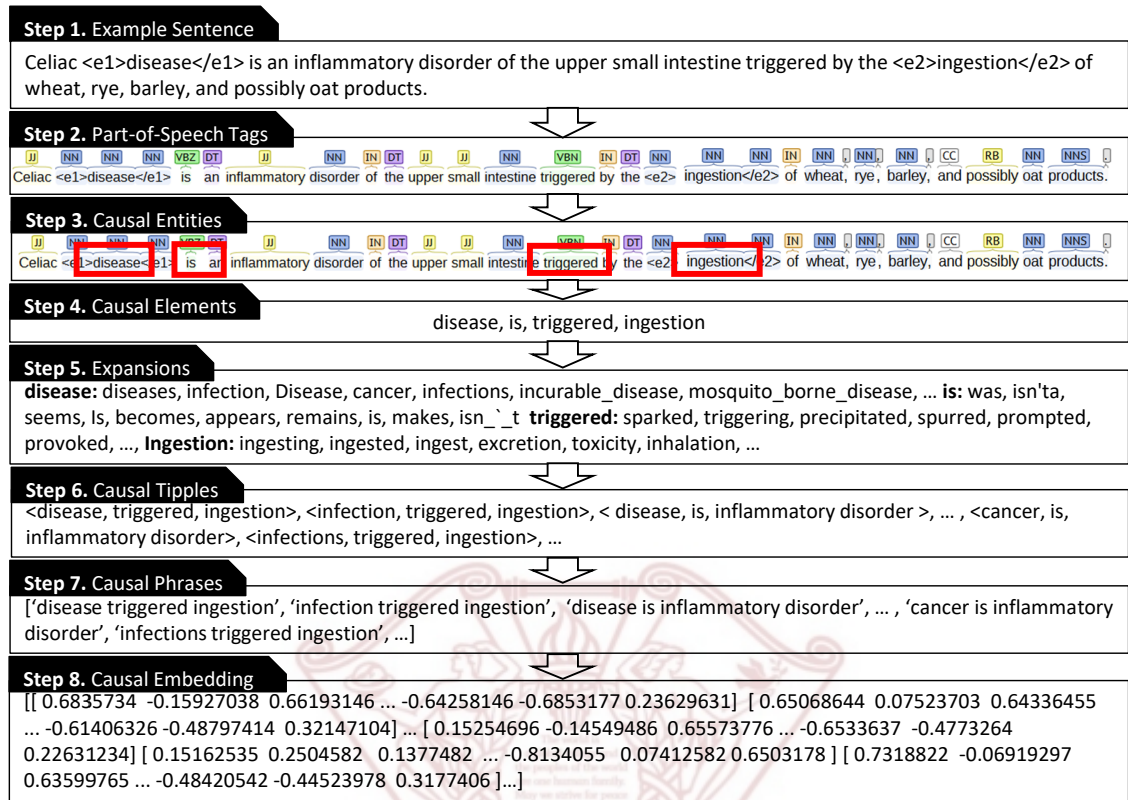


Figure 5.2: Training causal trigger extraction example

5.3 Causality Mining (CM) Module

The CM module is used for application of the CTTM on unseen, preprocessed test data, for classifying candidate phrases as causal or non-causal. This module utilizes three steps Candidate Triple Extraction, Causal Candidate Classification, and Triple Semantic Analysis, which are described in following sub-sections.

5.3.1 Candidate Triple Extraction

In the first step, starting with preprocessed sentences from unseen text, the Candidate Triple Extractor, identifies the candidate triples. These candidate triples are obtained by collecting all possible phrases of the form <NP, VP, NP> within each preprocessed sentence. This operation is performed in linear order to collect various candidate causal phrases within each sentence, thus

increasing the total number of candidates but greatly reducing the size of individual phrases. For sentences with more than one verb in a sentence, the noun phrases with longer dependencies are discarded. This is to maintain context of the nouns with their nearest verb phrase for matching with our causality identification patterns of SVO. An example of this process is shown in Figure 5.3, where the sentence from step 1, is pre-processed in step 2, before candidate triples for the same are generated in step 3. The candidate triples are then converted into candidate phrases (“*NP V NP*”), before the 6 BERT pre-trained models convert each of these into 6 embedded vectors.

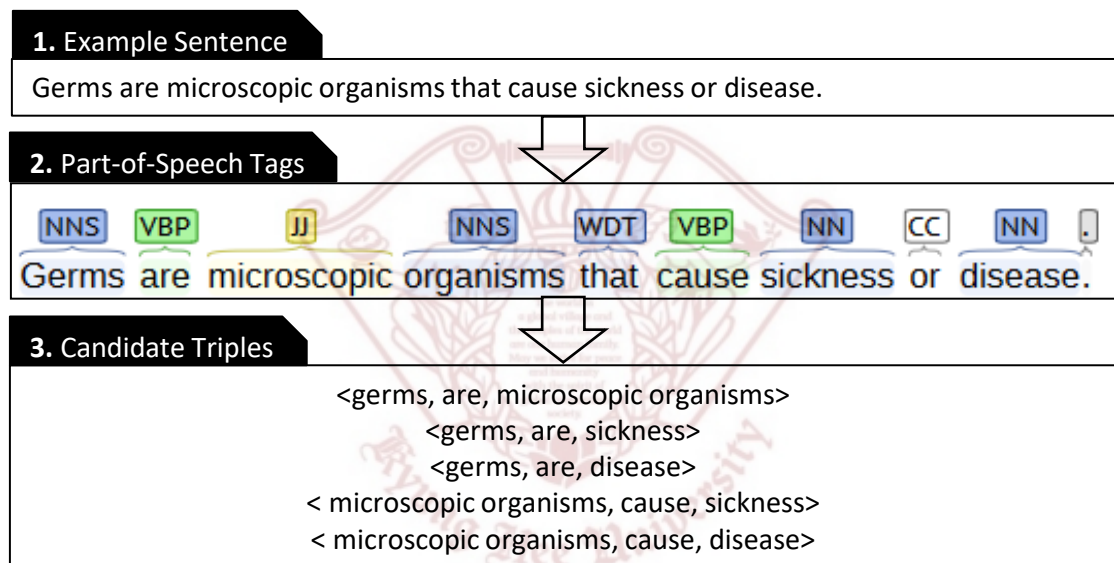


Figure 5.3: Training causal trigger extraction example

5.3.2 Causal Candidate Classification

Next, we apply the Causal Trigger Trained Model(CTTM) to classify the candidate embedded vectors generated in the previous step, as being causal or non-causal. The CTTM contains embedding vectors for 6 BERT models, which all participate in the causality classification operation, using cosine distance measure to solve this 2-class problem. Each of the BERT model, classifies a candidate triple as causal if the max similarity score is above α_i (where i is the index, corresponding to one of the six models, and α_i is computed using the threshold selection methodology

presented in 7.2.2). The causal triple thus identified is expanded by including the similarity score, as a fourth member, thus transforming the triple into a quad of form $\langle NP, VP, NP, [score_i] \rangle$. Where $score_i$, represents the similarity measure of a participating BERT model. These quads are then filtered using minimum similarity threshold. For 6 BERT models, presented in the Section 7.2, a candidate triple is thus classified as causal if at-least one model classifies it as causal. Additionally, the minimum value of $score_i$, greater than or equal to α_i is retained as the similarity score of the candidate triple. In this way, we can determine the minimum similarity of a candidate triple with most participating models. The final set of quad thus produced, pertains to causally classified instances only and is of the form $\langle NP, VP, NP, \min(score_i) \rangle$.

5.3.3 Triple Semantic Analyzer

The resulting set of quads, thus pertains to our classified positive class (causal) instances. While it may be possible to judge the classified instances, by extending the test data annotation of the sentence to the causal phrase, it is better to validate the classified instances from the expert. In order to support the expert, with maximum information about the classified instances (since conversion from corpus to sentence and then to candidate phrases removes a large part of their context), we extend each NP in the classified quad, with its associated Concept Unique Identifier (CUI) and semantic type using the UMLS REST API¹. This allows the system to identify if at-least one of the participating terms is semantically related to any medical terminology. If both terms do not have any corresponding concepts in UMLS, then it is also filtered out. The generation of this syntactically and semantically expanded set of classified instances then completes the process of lexical analysis and classification of the unseen clinical text.

5.4 Feedback Loop

A feedback loop allows the expert to validate the classified instances produced by the MD module by using the semantic information expanding the noun phrases of the causal quads and the similarity score. The expert can indicate a phrase as causal or non-causal, providing a basis for updating

¹<https://documentation.uts.nlm.nih.gov/rest/home.html>

the CTTM. This model evolution is achieved by generating the embedding vector for each of the expert validated causal classified instances, using BERT pre-trained models and simply appending the same to the training embedding vector list. Additionally, the phrases marked as non-causal, are added to a causal blocklist, which is then converted into an embedded vector, and compared with the vector lists of the CTTM. For each training embedding vector in the CTTM, if the similarity threshold with the blocklist embedded vector is greater than α_i , it is removed from the list. Initially, this lookup table is kept empty and as the expert identifies the correctness of causal phrases, it grows to include the correct phrases and discards similar non-causal phrases, for each of the six models. In this way, the CTTM evolves with each iteration and improves upon the previous results using expert feedback.

We validated the soundness of the proposed methodology by applying it on various datasets, and also compared the results with existing studies. As mentioned earlier, previous studies on causality classification have mainly focused on the creation and utilization of expert-generated rules. However, in a recent study [65], the authors presented a methodology, driven by the similarity between word embeddings, to classify instances from the same datasets we have used for evaluations. Their methodology is based on the identification of causal verbs between two labeled entities, followed by conversion of these verbs and those within the test data set into embedding vectors using Word2Vec. The embedding vectors are then compared using cosine similarity. If the similarity between the two vectors is greater than 0.5, the authors classify the verb from test instances as causal and add these verbs into the set of causal verbs used for subsequent matches. Finally using expert's rules, the authors classify the instance into one of the four different causality relationships (subject causes object, subject is the result of object, attribute relation, and certain relationship). The results presented by the authors indicate good performance of their model in comparison to two previous studies [54] and [57]. While the results presented by the authors in their manuscript are interesting, in their original form, they are incomparable to our results. We therefore, created an implementation of the Ning's strategy [65] to classify causal triples as causal or non-causal and compared the same with our results. During this implementation, we have utilized the same seed verb list, as presented by the authors in their research work, maintained the same similarity threshold value of 0.5, and followed the same design to classify each verb as

causal. For any triple, where the verb was classified as causal, the triple is also considered as causal. The results comparison is shown in Section 7.2.9.



The output triples of solution two, presented in Chapter 5, represent concepts connected through causality relationships. The direction of the causality can be either <cause, trigger, effect> or <effect, trigger, cause>. As our final goal is to produce production rules as for output. The rules follow IF condition(s) THEN conclusion format. The condition part represents cause(s), while the conclusion part represents the possible effect. Therefore, we can map the extracted triples to production rules as causal triples consist of both causes as well as their effects. However, in the production rule, a condition consists of three modules as key, operator, and value, while the extracted causal triples only consist of causal phrases which act as the key of the condition. This makes a gap between the extracted causal triples and the targeted output. This chapter deals with the identification of possible operator and value part of a rule to bridge the gap and produce a set of production rules.

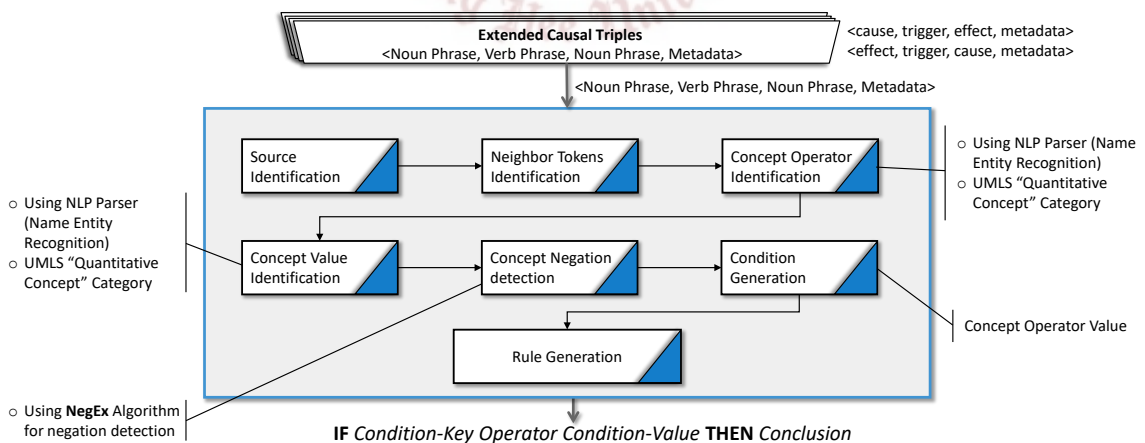


Figure 6.1: Rules generation workflow

6.1 Concepts Operator and Value Identification

To bridge the gap, we devise a methodology as shown in Figure 6.1 for causal concept operator and value identification. The proposed methodology is inspired by EXTEND a tool used for numerical value extraction from EMR data [68]. We evaluate the source sentences of the triples for both operator and value extraction. The steps required are described as follows.

- **Source Sentence Identification:** The context of a causal concept can provide the details including possible operator and value. The context can be built from the sourcing tokens, therefore, we keep track of the source sentence of each triple during the triple generation process. Each triple has its source context in the form of a complete sentence to be evaluated for an operator as well as value extraction.
- **Neighbor Tokens Identification:** A sentence can have multiple causal phrases, therefore we considered neighbor tokens for evaluation. This is to maintain the context of the causal noun with their nearest tokens. In the case where a sentence consists of only one causal phrase all tokens are considered for evaluation, while in other cases, only the nearest tokens with the causal phrase are considered as neighbor tokens.
- **Concept Operator Identification:** We apply two techniques for concepts' possible operator identification. First, we parse the source sentence with the Stanford NLP parser [75] to check the source sentence tokens POS tags, named entities, basic and enhanced dependencies. The evaluation enables us to identify and evaluate the relevant tokens surrounding the causal concept. The comparative token in the neighborhood of the causal concept is considered as the operator of the causal concept. Second, we identify the semantic type of each token of the source sentence from UMLS dictionary. The neighbor tokens having semantic type as "Quantitative Concept" are set as the possible operator. However if there is no comparative token we set "=" as the default operator for the causal concept.
- **Concept Value Identification:** The same procedure as mentioned in the above bullet point is followed for valued identification. However, our target here is to find the value, therefore, we focused on the "NUMBER" tag of the named entity recognition result of the Stanford

parser [75]. The neighbor “NUMBER” tagged token of a concept is set as the possible value of the causal concept. Similarly, from the UMLS semantic types, we considered “Quantitative Concept” tagged token as the possible value. However if we didn’t found any value we set the default value to “true”.

- **Concept Negation Detection:** The concepts mentioned in a clinical text may not always represent its presence, it can also reflect the absence if coupled with negation term. The negation identification is critically important in the clinical domain as it reflects the complete opposite meaning of the concept. Therefore, we adopt a widely used negation detection algorithm NegEx [88] for concept negation detection. A negated concept affects our identified operator in the condition. For example, a concept has operator “=” we detect that the concept is used in negation so the operator will be updated to “!=”. Similarly, the operator will be modified by adding “!” for negated concepts.
- **Condition Generation:** The causal concept, identified operator, and value are set into the production rule condition part. This will represent a condition as “concept operation value” which complete the rule condition.
- **Rule Generation:** The identified condition and associated effect concept are set into production rule format. However, each rule consists of only one condition and its effect because we transform each triple into a condition. While in reality, a rule may consist of multiple conditions. As mentioned earlier, a sentence can produce multiple triples, therefore, we combine rules having the same source sentences. This process reduced the number of rules but increase their effectiveness and make them applicable in the real field.

6.2 Example of Concepts Operator and Value Identification

The realization of the operator and value identification and extraction is demonstrated for an example triple in Figure 6.2. The example triple < Plasma glucose, diagnosed, diabetes> with the casual concept “plasma glucose” and effect concept “diabetes” is processed with the aforementioned steps to generate a production rule. The targeted triple is extracted from the “Patient with

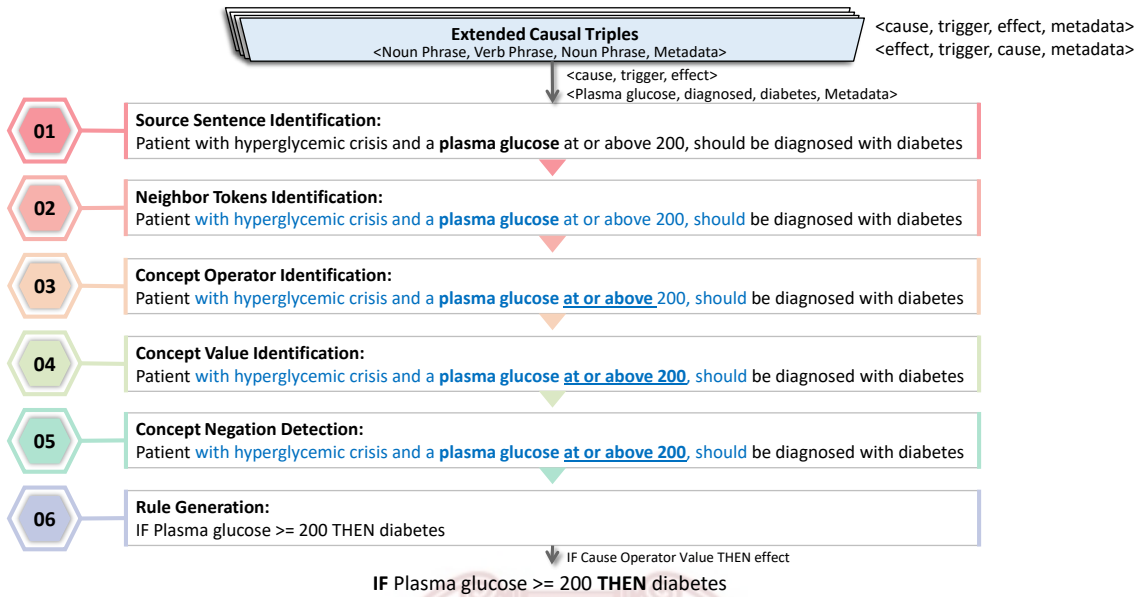


Figure 6.2: Example of rules generation from causal triples

hyperglycemic crisis and a plasma glucose at or above 200, should be diagnosed with diabetes” sentence, therefore, we need to evaluate this sentence for operator and value identification. The neighbor tokens window is set to five tokens. The tokens shown in red color in the Figure 6.2 are the neighbor tokens of the target concept “plasma glucose”. The source sentence is parsed with the Stanford NLP Parser [75] as shown in Figure 6.3. As we can see from the named entity recognition module of the parser, the number along with the operator is correctly identified in the source sentence. Therefore, we pick the operator and value for the rule condition. We also check the neighbors’ semantic type from UMLS for “Quantitative Concept” to verify the correctness of the identified values. The example sentence is also evaluated for “plasma glucose” negation. As the causal concept is not used in negation, therefore the operator remains the same as identified earlier. Finally, the target concept and identified operator and value are set into the condition of the rule, while the effect of the triple is set to the conclusion part of the rule. The rule “IF plasma glucose \geq 200 THEN diabetes” is generated as the final result for the triple.

The same process is repeated for all identified causal triples. As highlighted earlier, the resultant rules will have only one condition as shown in Figure 6.2. Rules acquired from triples with the same source sentences can be combined. Therefore, we combined the condition of the same

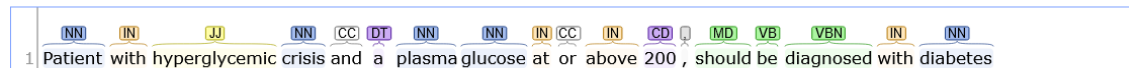
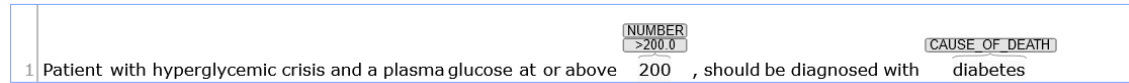
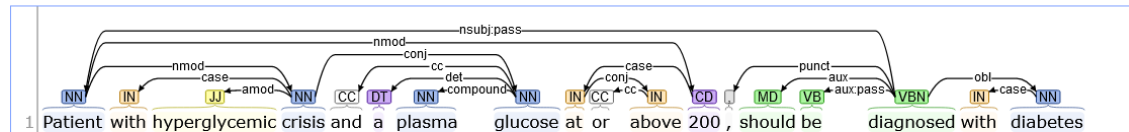
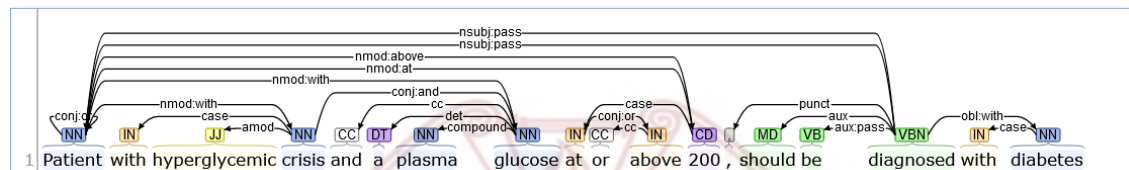
Part-of-Speech:**Named Entity Recognition:****Basic Dependencies:****Enhanced++ Dependencies:**

Figure 6.3: Example sentence parsed via the Stanford parser

sourced sentence rules with a similar conclusion. The resultant rules with multiple conditions can better assist in clinical decision-making as clinicians prefer to evaluate multiple conditions before taking any clinical decision.

6.3 An End-to-End Example

The end-to-end realization and techniques used at each step of the proposed methodology is shown in Figure 6.4. The methodology takes clinical document(s) as input and processes it through six major steps for knowledge acquisition. Step one splits the input documents into sentences via NLTK sentence tokenizer. Step two evaluates each sentence and tags it as a recommendation or non-recommendation sentence based on the presented information of the sentence. This step uses one of the classification method presented in the Chapter 4. The recommendation tagged sentences are further processed for knowledge acquisition. We processed the example recommendation sentence “Patient with hyperglycemia or hyperglycemic crisis, and a plasma glucose ≥ 200 , diagnosed with diabetes” to clarify the input and out of each step involved in the knowledge

The diagram illustrates a comprehensive NLP-based framework for clinical text analysis, organized into several key components and processes:

- Clinical Docs**: The source of clinical text.
- Sentence Extractor**: Extracts sentences from clinical documents.
- Sentence Classification**: Classifies sentences based on recommendation sentences (e.g., "Patient with hyperglycemia or hyperglycemic crisis, and a plasma glucose >= 200, diagnosed with diabetes").
- Text Preprocessing**:
 - Tokenizer**: Converts sentences into tokenized sentences (e.g., ["Patient", "with", "hyperglycemia", "or", "hyperglycemic", "crisis", "..."]).
 - POS Tagger**: Assigns Part-of-Speech (POS) tags to the tokenized sentences (e.g., ["Patient", "NN", "with", "NNP", "NNPS", "POS", "Tag", "with", "neighbor", "JJ", "..."]).
- Components Output**: A table listing the techniques used in the framework:

No.	Techniques
1	NLTK Sentence Tokenizer
2	Pattern based OR ML based OR Deep Learning based classification
3	NLTK word Tokenizer
4	NLTK POS Tagger
5	Concenate triple by space
6	BERT NLI models
7	Cosine Similarity based multi-model classification
8	Cause phrase with nearest Quantitative terms
9	Effect phrase
10	IF Cause(s) THEN Effects
- Triple Generation**:
 - Noun Extraction**: Extracts nouns from sentences (e.g., ["Patient", "NN", "with", "NN", "hyperglycemia", "NN", "or", "CC", "hyperglycemic", "JJ", "crisis", "NN", "..."]).
 - Verb Extraction**: Extracts verbs from sentences (e.g., ["Patient", "NN", "with", "NN", "hyperglycemia", "NN", "or", "CC", "hyperglycemic", "JJ", "crisis", "NN", "..."]).
 - Triple Generation**: Generates triples from the extracted nouns and verbs (e.g., ["Patient", "NN", "with", "NN", "hyperglycemia", "NN", "or", "CC", "hyperglycemic", "JJ", "crisis", "NN", "..."]).
- Semantic Analyzer**:
 - Triple classification**: Classifies triples into causal triples (e.g., ["hyperglycemia", "diagnosed", "diabetes", "hyperglycemic", "crisis", "diagnosed", "diabetes", "plasma", "glucose", "diagnosed", "diabetes"]).
 - Triple embedding**: Converts triples into numerical embeddings (e.g., [0.07387608, 0.23522191, 0.65920086, ...]).
 - Triple phase generator**: Generates triple phases (e.g., ["Patient", "diagnosed", "diabetes", "hyperglycemic", "crisis", "diagnosed", "diabetes", "plasma", "glucose", "diagnosed", "diabetes"]).
- Knowledge Creation**:
 - Conclusion Identifier**: Identifies conclusions from triples.
 - Condition Identifier**: Identifies conditions from triples.
 - Rule Generator**: Generates rules based on the identified conclusions and conditions (e.g., "IF hyperglycemia = Yes AND hyperglycemic crisis = Yes AND plasma glucose >= 200 THEN Diabetes").
- Knowledge base**: A database storing the generated rules and triples for future use.

Step three performs text pre-processing on the recommendation sentence. The pre-processing includes tokenization which breaks down the sentence into tokens, and POS tagging where each token is tagged with its most appropriate POS tag using NLTK POS Tagger. This produced POS-tagged tokens of the sentence as output. Step four aims to generate triples of the form < Noun Phrase, Verb Phrase, Noun Phrase> out of the POS tagged sentence. Therefore, it looks at the POS tags of each token to locate noun and verb-tagged tokens. The sentence tokens tagged with “NN”, “NNS”, “NNP”, or “NNPS” are considered noun phrases in our study. Similarly, “VB”, “VBD”, “VBG”, “VBN”, “VBP”, or “VBZ” tagged tokens are considered verb phrase of the sentence. All possible combinations of the verb phrase encapsulated by the noun phrases are captured as candidate triples. The aforementioned example sentence produced four candidate triples as shown in step four of the Figure 6.4. The generated candidate triples are produced as output of this step and passed to subsequent steps for further processing.

Collection @ khu

transformed into embedding vectors using six BERT models. The generated embeddings are compared with each model's corresponding causal triple embeddings to measure the similarity of the triple. Each model tags the triple as causal if its similarity is greater than a threshold value, non-causal otherwise. Finally, a triple tagged as casual by at least one BERT model is considered as causal triple. Among the four candidate triples produced from the example sentence, three triples were tagged as causal by at least one model while there is only one triple "<Patient, diagnosed, diabetes>" classified as non-causal by all six BERT models. As we can see the identified non-causal triple has no causal relation between the concept "patient" and "diabetes" which indicates our models are able to correctly distinguish between causal and non-causal concepts.

Finally, the extracted casual triples are evaluated for rule generation at step six of the methodology. As all three triples follow < cause, trigger, effect> format. Therefore, the first concept (cause concept) is set as the rule condition key while the second concept (effect) is set as the conclusion of the rule. The source sentence is evaluated to find the operator and value of the causal concept. In the example sentence two of the causal concepts "hyperglycemia", and "Hyperglycemic crisis" are having no quantitative concept and negation clue in their neighborhood therefore, their operators are set to "=" and values as "Yes". The third concept "plasma glucose" has comparative token ">=" as well as quantitative token "200" therefore, its operator and value is to ">=", and "200", respectively. Since all the triples are generated from a single sentence, and the conclusion part of all three triples are same as "diabetes". Therefore, we combined all three rules into a single rule as "IF hyperglycemia = Yes AND hyperglycemic crisis = Yes AND plasma glucose >= 200 THEN Diabetes". The extracted rule is stored into the knowledge base to be used by human experts or automated healthcare systems for better clinical decisions.

As described, the proposed methodology consists of three main parts/solutions, therefore, this chapter provides the results and evaluation of each solution in the following sections.

7.1 Text Classification Results

7.1.1 Text Classification Dataset

We evaluated the proposed sentence classification methodology based on the system's accuracy. The dataset used for this module consists of three guidelines including Hypertension [78], Rhinosinusitis [89], and chapter 4 of the asthma guideline [90]. Each sentence of the guidelines is annotated by physician as Condition-Action (CA), Condition Consequences (CC), Action (A), or Not Applicable (NA). However, we considered CA, CC, and A tagged as recommendation sentences RS while NA tagged sentences as NRS. The expert provided label of each sentence is considered a ground truth label and compared with system generated label for evaluating the accuracy of the methodology. As shown in the Table 7.1, the hypertension guideline consists of 78 recommendation sentences out of 278 sentences, rhinosinusitis contains of 151 recommendation sentences among 761, and Asthma has total 53 recommendation and 118 non-recommendation sentences.

Table 7.1: Details of text classification dataset

Guideline	Total Sentence	Recommendation Sentences	Non-Recommendation Sentences
Hypertension	278	78 (28.06%)	200 (71.94%)
Rhinosinusitis	761	151 (19.84%)	610 (80.16%)
Asthma	171	53 (30.99%)	118 (69.01%)

7.1.2 Pattern based Classification Results

We used 70% of the hypertension guideline for manual pattern extraction and machine learning based salient term extraction, while the remaining 30% of the hypertension guideline were used for pattern evaluation. Furthermore, we evaluated the extracted patterns on Rhinosinusitis [89] and chapter 4 of asthma guideline [90] to check the generalization and accuracy of the extracted patterns. The manual as well machine based processing of the CPGs required text preprocessing to clean and prepare the CPG content. The preprocessing steps required for KEs were simple, and the only requirement was to split the CPG documents into sentences. However, the preprocessing steps required for machine learning models are more impactful in terms of the final model accuracy and the number of salient terms. We compared the models with applying feature selection techniques and without feature selection. We used the information gain ratio to assign a weight to features and selected top k features. As mentioned earlier, the value of k highly affects the model accuracy and the salient terms considered by the model. Therefore, we tested the model on different values of k . The detail of k values and their effects on the accuracy of the decision tree model is shown in Figure 7.1. As shown in Figure 7.1, initially the accuracy was increasing gradually with an increment of k value. From $k = 40$ to $k = 79$ the accuracy remained stable with maximum value, while accuracy started to decrease as the value of k increased from 79. The accuracy of the decision tree model in maximum at $k = 40$. Therefore, we selected top 40 features for model training i.e. $k = 40$. The accuracy starts decreasing due to less relevant terms consideration as k approaches beyond 79.

To extract salient term from the trained white box machine learning models, we evaluated our trained models: decision tree, rule induction, and gradient boosted tree with and without feature selection on the hypertension [78], rhinosinusitis [89] and chapter 4 of asthma guideline [90]. The models achieved classification accuracy as given in Figure 7.2. Where graph (a) represents model accuracies when features selection was not performed and (b) represents accuracies with features selection. Based on the results shown in Figure 7.2, the accuracy of the model increases with feature selection. Also, the final generated model changes the extracted salient terms.

As described, we have three types of patterns; heuristics patterns, POS based patterns, and UMLS based patterns. The CPGs sentence classification accuracy of each approach is given in the

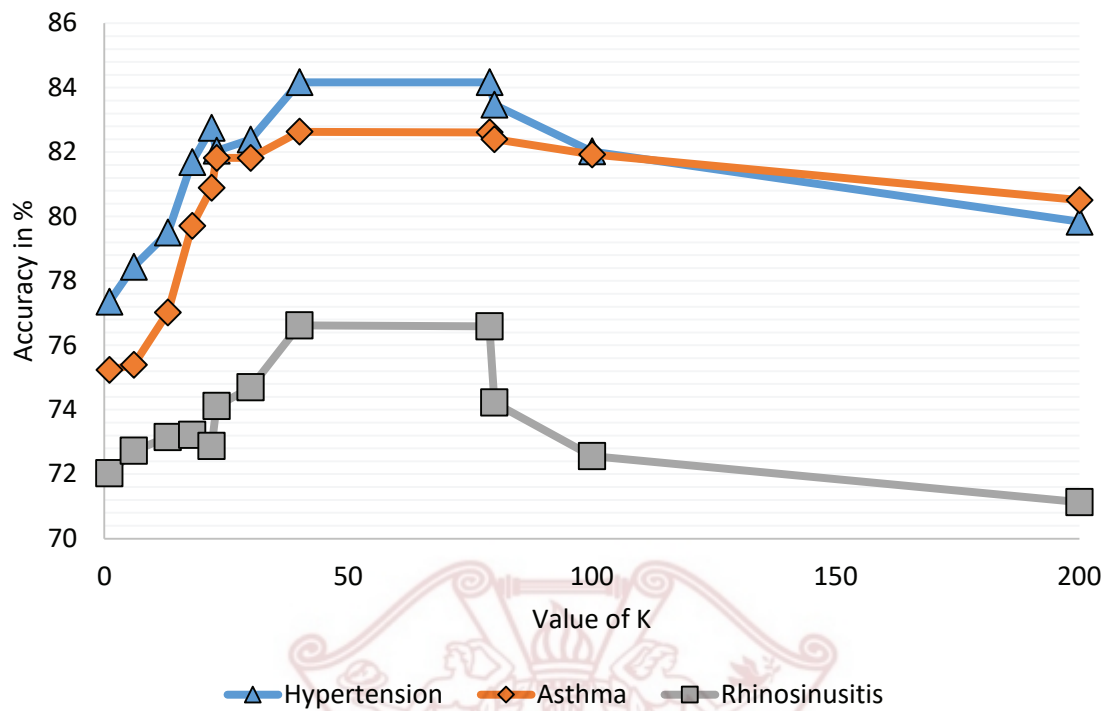


Figure 7.1: Top k features and the model accuracy

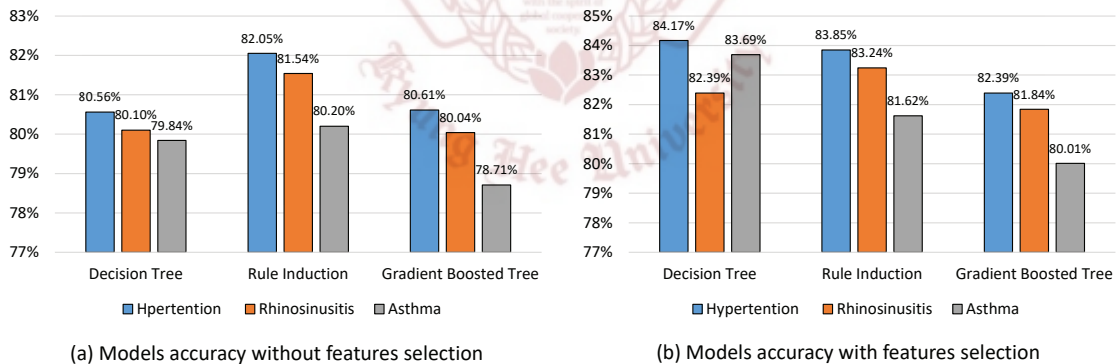


Figure 7.2: Models accuracy without and with features selection

subsequent subsections.

Heuristic Patterns Results

The heuristic pattern-based method without considering the salient terms list gives 84.93% accuracy on the test dataset (30% of the hypertension guideline). The results showed that the extracted

patterns work well on the test dataset. The extracted patterns, given in Table 4.2 of chapter 4, were also applied on Rhinosinusitis [89] and chapter 4 of asthma [90] guidelines to evaluate the accuracy of the extracted patterns. Our proposed method achieved an accuracy of 71.93%, 75.56%, and 84.93% on asthma [90], Rhinosinusitis [89], and Hypertension [78], guidelines, respectively, as depicted in Figure 7.3(a). When the patterns were reevaluated by considering machine learning extracted salient terms, KEs updated the patterns as shown in Table 4.4 of chapter 4 that result increase in accuracy to 73.29%, 74.37%, and 86.04% in asthma [90], Rhinosinusitis [89], and Hypertension [78], guidelines, respectively as shown in Figure 7.3(b).

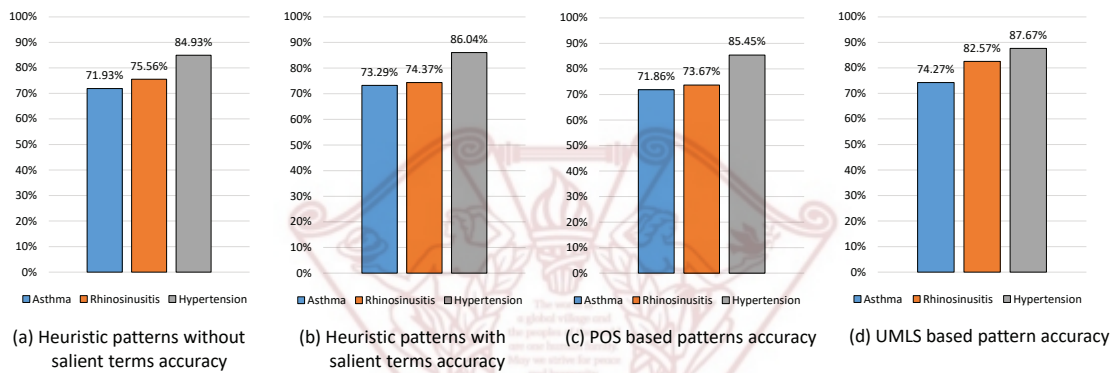


Figure 7.3: Extracted patterns accuracy

The heuristic patterns performed well on the testing part (remaining 30%) of the hypertension guideline [78]. However, the accuracy decreased by 12.75% on the other two guidelines i.e., asthma and rhinosinusitis. The primary reason for this low accuracy was the diverse format of the guidelines. One CPG uses different words and their sequence for representing the same concepts as the others. Therefore, to overcome this issue and to maintain accuracy, we added the POS based patterns into the proposed technique.

POS Patterns Results

In the POS based pattern technique, we combined the POS tags with clue words of the RS sentences. Because the combination of POS tags and the clue words increased the system accuracy. To evaluate the accuracy of the technique, all three guidelines (asthma, rhinosinusitis, and hypertension) were used in the experiment, and we achieved an accuracy of 71.86%, 73.67%, and

85.45%, respectively, as shown in Figure 7.3(c).

The results of Figure 7.3(c) depicts that the POS-based pattern did not perform well than the heuristic patterns. However, POS patterns are applicable on all CPGs irrespective of the CPG format. We achieved better accuracy than the POS without clue words, the primary reason was the generalization of the patterns along with clue words. However, some of the clue words may not be used in different guidelines. Therefore, a complete and generic solution is required to resolve the aforementioned problem. To remove this deficiency, we merged UMLS based patterns into the proposed technique, which increased the system accuracy. The detailed results of the UMLS pattern are described in the following subsection.

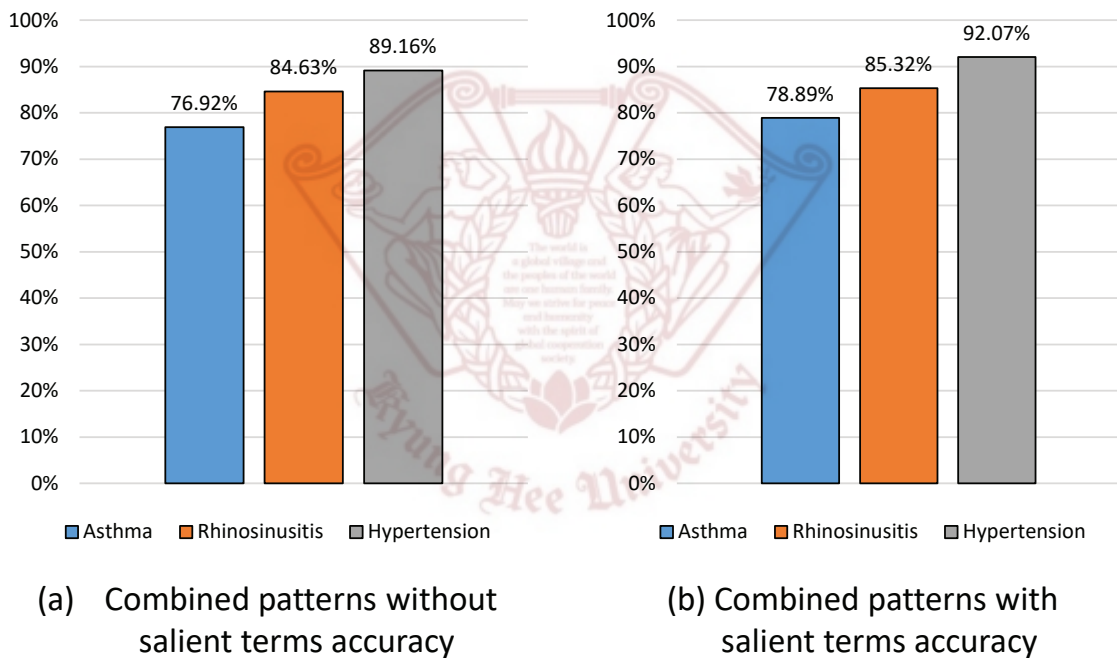


Figure 7.4: Combined patterns accuracy with and without salient terms.

UMLS Patterns

The UMLS patterns, given in Table 4.7, classified recommendation sentences with the accuracy of 74.27%, 82.57%, and 87.67% for asthma, rhinosinusitis, and hypertension guidelines, respectively, as shown in Figure 7.3(d). The reason for the improvement of accuracy was the UMLS concepts used in the recommendation sentences. Mostly, the recommendation sentences use tags

of "Population Group", and "Pharmacologic Substance"; therefore, UMLS based patterns can easily recognize these sentences and increase the accuracy of the systems' classification.

After individual evaluation, we combined all three techniques and evaluated asthma, Rhinosinusitis, and Hypertension guidelines before providing salient terms and after providing salient terms. Before using salient terms the extracted patterns achieved the accuracy of 76.92%, 84.63%, and 89.16%, respectively, as shown in Figure 7.4(a). However, after using salient terms the patterns accuracy increased to 78.89%, 85.32%, and 92.07%, respectively, as shown in Figure 7.4(b). Here each sentence was evaluated by the three types of patterns and tagged independently. A sentence tagged by one or more techniques was finally considered as an RS sentence otherwise NRS.

As shown in Figure 7.2, 7.3, and 7.4 the feature selection, salient terms, and combined patterns increased the classification accuracy, respectively. However, we performed a non-parametric p -value test to check the significance of the improvements [91]. The improvement shown in Figure 7.2 via feature selection (hereafter Model FS) compared to without feature selection (hereafter Model WFS) is evaluated with a threshold value of .05 under the following hypothesis.

- Null hypothesis H_0 : Model FS isn't better than Model WFS
- Alternate hypothesis H_1 : FS is better than WFS

The calculated p -value for the above hypothesis is .035, which is less than the threshold value of .05. Therefore, it rejects the null hypothesis H_0 and conclude that model FS is better than WFS. Similarly, we calculated the p -value for other two cases, with and without salient terms 7.3, and combined vs individual patterns 7.4 with resulted value of .038 and .040, respectively. Hence the p -values showed the improvement caused by feature selection, salient terms, and combination of heuristics, POS, and UMLS patterns are statistically significant.

Text Classification Evaluation

The proposed technique is evaluated and compared with existing classical and advanced machine learning models. In classical models, we targeted zeroR, Naive Bayes, J48, and Random Forest as shown in 7.5 (a), while in advanced models, our focused algorithms are neural network (CNN),

long short-term memory (LSTM) and Bi-directional LSTM (Bi-LSTM) as shown in 7.5 (b). In the deep learning models, inspired from [26, 92], we used embedding size to 1776, adam optimizer, binary cross-entropy as loss function, and set dropout value to 0.5. In classical models, ZeroR achieved 69%, Naive Bayes 69%, J48 67%, and Random Forest achieved an accuracy of 67% on asthma guideline, however, the proposed approach achieved higher accuracy of 78.89%. Similarly, the accuracies of these algorithms on Rhinosinusitis guideline were, 80%, 80%, 81%, 84%, respectively, while the proposed technique performed better with accuracy of 85.32%. Likewise, the proposed algorithm correctly classified Hypertension CPG sentence with an accuracy of 90.07%, which is higher than all classical models as depicted in 7.5 (a). The improved results of the proposed methodology are mainly due to the relevant patterns execution, by combining expert heuristics with machine learning techniques, and the generalization of the patterns through POS, and UMLS techniques.

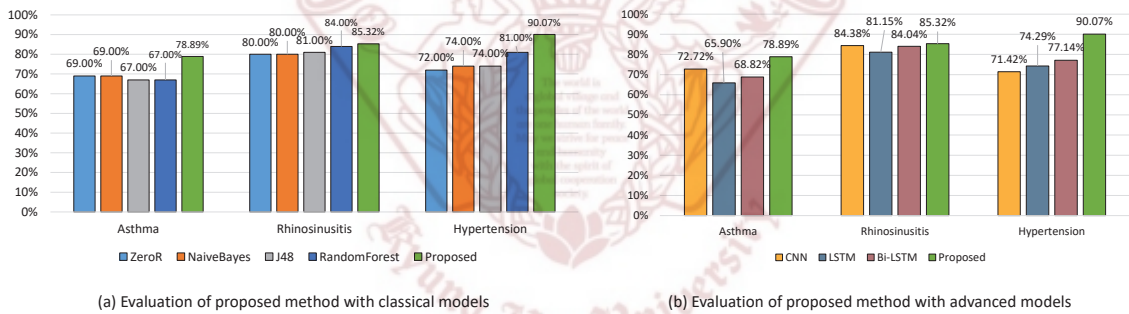


Figure 7.5: Extracted patterns evaluation

In advanced models, the accuracy of CNN is 72.72%, LSTM is 65.90%, Bi-LSTM is 68.82%, and the proposed system is 78.89% on asthma guideline. On Rhinosinusitis CPG, the accuracies were 84.38%, 81.15%, 84.04%, and 85.32%, respectively. While, in the Hypertension guideline, our proposed approach showed better results than the advance machine learning models, which is 90.07% higher than 71.42%, 74.29, and 77.14% as shown in 7.5 (b). The results obtained from the deep learning models surpassed the classical models in terms of accuracy. However, the proposed technique performed better than deep learning models. This is mainly due to the fact that deep learning models are data hungry models and required a large training data than the provided one.

7.1.3 Traditional Machine Learning based Classification Results

We evaluate, the application of machine learning for clinical text classification using Rapid Miner studio [93]. The algorithms evaluated for the task includes Naive Bayes, Generalized Linear Model, Deep learning (a shallow model), Decision Tree, Random Forest, and their ensemble. We already compared some of these models with our pattern based approach as shown in Figure 7.5. However, the major concern here is to explore the effect of feature extension on the performance of these models as discussed in Section 4.2. We performed multiple experiments with different settings on annotated hypertension CPG [78] consists of 78 recommendation statements among total 278 statements. The CPG was split into 70% and 30% for training and testing part. The training part of the CPG consist of total 195 statements including 58 recommendation statements. While the testing part consists of total 83 statements including 20 recommendation statements. The trained models were also validated on Rhinosinusitis CPG [89] to authenticate the performance (in term of accuracy) of the models.

The experiment that outperformed among others achieved the best accuracy of 79.82% by Ensemble Learner algorithm as shown in Figure 7.6(a). In this experiment, we used TF-IDF for word vector generation, Non Letters for tokenization, WordNet for stemming and English stopwords were filtered out. In the filter tokens component, we observed from multiple experiments that the NN and NB tokens have the maximum contribution in achieving the accurate result. Therefore, we filter out all other tokens. We find the synonyms of the remaining token using WordNet dictionary. However, in this experiment the aspects/concepts for input dataset were not included.

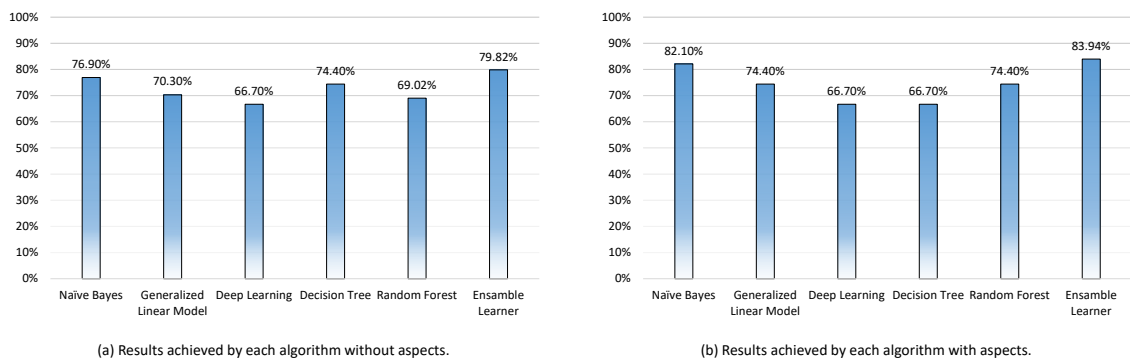


Figure 7.6: Machine Learning Models Classification Results

The experiment was repeated with the same setting as earlier but we also find and include aspects because clinical guidelines describe clinical scenarios and normally uses clinical terminology specific to a target disease. To extend the scope of the mechanism to be applicable on any CPG irrespective of the target disease we find the category (Aspect/Concepts) by utilizing UMLS medical dictionary. The final structured data generated is consists of word tokens, their synonyms and their aspect along with occurrence frequency. We trained and tested machine learning models. The models considered for the study includes Naïve Bayes, Generalized Liner Model, Deep Learning, Decision Tree, Random Forest, and Ensemble Learner as shown in Figure 7.6(b). The models achieved 82.10%, 74.40%, 66.70%, 66.70%, 74.40%, and 83.94% accuracies respectively.

7.1.4 Deep Learning based Classification Results

As mentioned, the datasets used for training (70% of the hypertension guideline) have a small number of sentences, and the distribution between recommendation and non-recommendation sentences is also very biased towards non-recommendation. Therefore, data-hungry models like deep learning models did not perform well as shown in 7.5 (b). To overcome this deficiency, We checked the applications of these advanced models with a large datasets by bootstrapping our dataset. Three different experiments using bootstrapping and data balancing techniques were performed and the results obtained are shown in Figure 7.7.

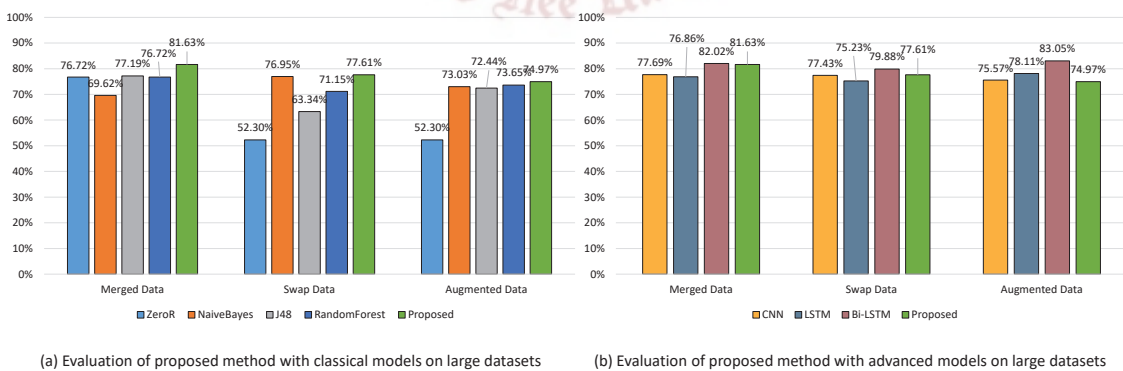


Figure 7.7: Evaluation of proposed method on large datasets

Initially, we merged all three datasets given in Table 7.1 resulted in a comparatively large and an imbalanced dataset of 1210 sentences with 282 recommendation and 928 non-recommendation

sentences. We named the generated dataset as "Merged Data". The application of classical and advanced machine learning models on this dataset is shown in Figure 7.7 (a) and (b), respectively. Among the classical model, decision tree (J48) model performed the best at an accuracy of 77.19%, but still bellow the proposed technique which stands at 81.63%. While in deep learning models CNN achieved 77.69%, LSTM 76.86%, and Bi-LSTM surpassed the proposed technique by 0.39%. The merged dataset is more inclined toward non-recommendation sentences, therefore, the trained models are also biased toward the non-recommendation sentence. We overcome dataset biases by duplicating the number of RS sentences, and swap their tokens, repeatedly. The resultant dataset referred to as "Swap Data" in Figure 7.7 consist of 846 RS and 929 NRS of 1775 sentences. The evaluation results of classical and deep learning models on Swap Data are reflected in Figure 7.7, where the Naive Bayes achieved the highest accuracy of 76.95% in classical models while Bi-LSTM achieved highest accuracy of 79.88% in deep learning model compared to 77.61% accuracy of the proposed technique.

Duplicating instances and swapping tokens may not be an efficient approach for trained a generalized model. Therefore, we balanced and enlarge the dataset by data augmentation [83], where we generated various RS sentences from the existing RS sentences by replacing word tokens with their synonyms. The resultant dataset referred to as "Augmented Data" in Figure 7.7 consists of 846 RS, 929 NRS sentences. The application of classical and deep learning models on the augmented data is shown in Figure Figure 7.7 where the naive based remains at top , however its accuracy dropped to 73.03%, while the proposed method accuracy dropped to 74.97% highest in the classical models. Similar to the previous cases, Bi-LSTM remains at top by achieving an accuracy of 83.05%, 8.08% higher than the proposed technique. Despite better performance of deep learning models, the tree based and pattern based approaches are preferred in real clinical practices. Because the pattern based approaches performs well on small datasets compared to deep learning models as observed from results in Figure 7.5 (b). Additionally, clinical decision making needs transparent solutions to enhance the physicians satisfaction. However, the pattern based decision making is traceable instead of deep learning models.

7.2 Causality Mining Results

7.2.1 Experimental Setup

The causality mining methodology presented in Chapter 5, represents a theoretical framework for identifying causal relationships in unstructured text. In order to build a sound realization of this framework, it is pertinent to identify the concrete models and algorithms, which can locally optimize each component, providing intermediate results with high performance and in turn amalgamate the workflows, providing a global optimal result for causality mining. Through various experiments we evaluated the impact of causal term expansion models, embedded vector generation methodologies, and similarity thresholds calculation to identify a well-balanced ecosystem, fulfilling our local and global optimization objectives. The experimental setup can be categorized into 3 stages, as shown in Figure 7.8, where each following stage, receives data from all previous stages.

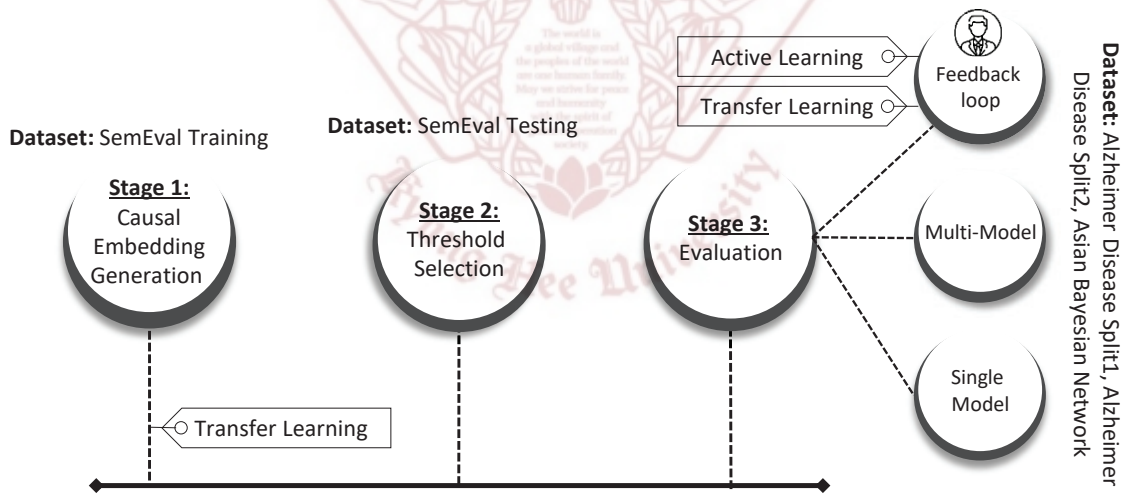


Figure 7.8: Causality mining experimental setups

Stage 1 - Causal Embedding Generation

In Stage 1, Causal Embeddings were generated for the SemEval 2010 task 8 training dataset [94], using the six pre-trained BERT models. This dataset pertains to the semantic relation identification process and identifies the relationships between nominals for drug-drug interactions from

biomedical texts. Each sentence in this training dataset and its counter part SemEval 2010 task 8 test dataset [94], is tagged with its most plausible truth-conditional interpretation using one of product-producer, content-container, cause-effect, and other semantic relations. However, since the target of this study is causality mining, we therefore, only considered the cause-effect tag as casual relation and all other as non-causal relations. The SemEval 2010 task 8 training dataset [94] comprises of 1003 causal sentences out of 8000 sentences. From these 1003 causal sentences, we extracted 1071 unique causal triples. The verb within each triple is then expanded using the pre-trained Google News model [37]. After the expansion, we take Cartesian product of the two encapsulating nouns of the source triple and one of the expanded verb to produce a little over 1.2 million expanded triples. Thus, with this expansion we are able to classify a wider range of causal relations, than what would have been possible, otherwise.

Next we convert these expanded triples into embedding vectors using six pre-trained BERT NLI models [87, 95], which include nli-base-mean-tokens, nli-large-mean-tokens, nli-base-max-tokens, nli-large-max-tokens, nli-base-cls-token, and nli-large-cls-token. These model differ in terms of their size (base or large) and the pooling layer used at the end of their deep neural network (mean pooling word tokens, max pooling word tokens, or cls pooling sentence token). Embedding vector generation for the 1.2 million expanded triples is a computationally expensive operation, which can take several days running on the CPU, however, due to the ability of the sentence_transformer library in python, to optimally use GPU, if available, the computational time is reduced, substantially. Through our experiments, we were able to process the expanded triples and produce the embedding vectors for base models in under 20 mins each and for large models in an hour, each. Overall, the embedding vectors were produced in 4 hours, using NVIDIA GeForce RTX 2060 GPU.

Stage 2 - Threshold Selection

Stage 2 is designed for threshold selection, whereby a sentence can be categorized as causal or non-causal, based on its similarity with the expanded triple set. Similarity threshold plays a vital role in the causality classification process and therefore requires extensive experimentation to select the best similarity score, above which a triple can be classified as causal. In order to fulfill

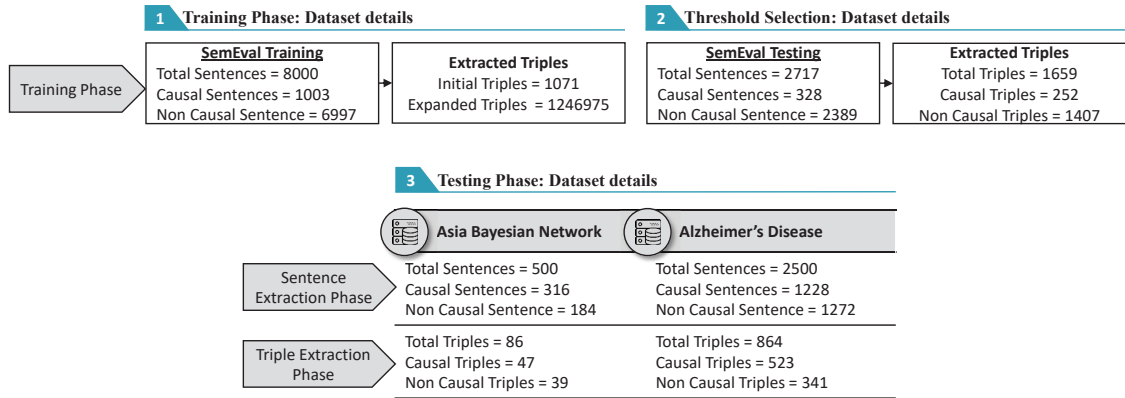


Figure 7.9: Details of causality mining dataset.

this aim, we utilized SemEval 2010 Task 8 test dataset to learn the best threshold value, where the precision-recall curve (PRC) obtains maximum area under the curve. In biased datasets, where the ratio of positive class is much lower than the negative class, Area under the PRC (AUPRC) is an optimal metric for selecting the threshold [96]. As shown in Figure 7.9, the SemEval 2010 Task 8 test dataset [94], contains 328 causal sentences, out of a 2717 total sentences (12.07% of positive class). We utilized AUPRC to learn optimal threshold values for each BERT models. The detailed result of threshold selection will be presented in Section 7.2.2.

Stage 3 - Evaluation

In Stage 3, we performed single model, multi-model, and multi-model with feedback loop evaluations on the Asian Bayesian Network dataset [65] and the risk factors of Alzheimer's disease (AD) [65], using the causal embedding vectors from Stage 1 and threshold values from Stage 2. The AD dataset consists of 1228 causal sentences out of 2500 sentences, while the Asian Bayesian Network dataset have 316 causal sentences from a set of 500 sentences. The sentences in these two datasets are tagged with either NP→NP (Noun Phrase influences Noun Phrase), NP-NP (Noun Phrase is related to Noun Phrase), or NP×NP (both nouns are irrelevant) label. In this study, we considered the first two tags (NP → NP and NP-NP) as causal and the remaining (NP×NP) as non-causal. Due to the large size of AD dataset and to test various iterations of the feedback loop, we split this dataset into two parts, using random selection for 50% partitioning. The complete

AD dataset, contains 864 candidate triples, out of which 523 are causal (60.53%) and 332 are non-causal (39.47%). With 50% random split, the AD1 and AD2 dataset contain 432 triples each. AD1 contains 267 actual causal triples(61.80%) and AD2 contains 256 actual causal triples (59.26%). Evaluations by all three methodologies (single model, multi-model, and multi-model with feedback loop) were performed on these three instances of the datasets (AD1, AD2, and Asian Bayesian Network). This data split is especially, important to execute and evaluate multiple iterations of the feedback loop, on unseen data.

In single model, we evaluate the performance of each BERT model to check the effect of the model size in terms of base and large, and pooling strategies using CLS-token, mean of all output vectors, and max-over-time of the output vectors and select a single best performing model for causality mining. However, by inspecting the result of each BERT model in terms of unique causal triple identification via a very handy UpSet tool [97], which can plot associations between different sets and can be used to visualize relationships, where the traditional Venn diagrams may fail (such as when the number of sets are greater than 4)¹. Since the aim of our approach is to improve the accuracy of causal classification, even in presence of false positives, it is then pertinent to analyze the UpSet results, based on a “minimum” intersection degree metric. This entails, the evaluation of causal classifications for a minimum intersection degree such as degree ≥ 1 , degree ≥ 2 and above. Intuitively, it can be seen that the performance results for degree ≥ 2 should be less than the performance for degree ≥ 1 and leads to a multi-model evaluation. The UpSet analysis performed in Section 7.2.7 revealed to used multi-model evaluation to increase efficiently of the causality mining.

In multi-model evaluations, we performed the experiments on the same three test datasets. However, in this case, we considered a triple as causal if any of the six BERT models tagged it as causal and non-causal otherwise. The results achieved in multi-model evaluation is shown in subsection 7.2.7.

Finally, we incorporated human expert’s feedback into the multi-model similarity matching process, to analyze the change in the quality of causality detection. For this process, an expert

¹The interactive UI is available at <http://vcg.github.io/upset/?dataset=10>, with the data drescription file for our presented approaches present at <https://raw.githubusercontent.com/Musarratpcr/CausalityDetection/master/Revision1/ADandAsianDatasetUpsetDescription.json>

(physician) from our collaborative hospital, verified the accuracy of the classified sentences. Since our automated process is dependent upon various datasets and has been repurposed, as explained earlier, this secondary verification is of utmost importance. This process was repeated in three iterations, while we ensured that once the CTTM is updated by the embedding vector of an expert verified causal triple, the same is not made a part of any subsequent test sets. Thus, the test sets in each iteration remain unseen. In Iteration-1, we used the embedded models (CTTM) trained on the SemEval 2010 Task 8 training dataset, and tested using the AD1 dataset. Embedding vectors corresponding to the correctly classified and expert verified causal and non-causal triples were then used to update the CTTM. In Iteration-2, this updated CTTM was then used to test the candidate triples from AD2 dataset. Once again, the correctly classified and expert verified causal and non-causal triples were used to again update the CTTM. Finally in iteration 3, the most recently updated version of the CTTM was then used for classifying the candidate triples from the Asia dataset. The details of the results achieved in each iteration are described in Section 7.2.8.

For experimentation, we used python code on Google Colab, with many additional libraries including Gensim models, NLTK, BERT sentence.tranformer, and sklearn. Using the same settings we developed a python based end-to-end application, which can extract causal relationships from an unseen copora. The application of the causality mining methodology and its evaluation was run on a dedicated workstation with Intel(R) Core(TM) i9-9900KF CPU, with 64GB ram, and NVIDIA GeForce RTX 2060 GPU. The training model was produced in under 4 hours, using a combination of CPU(for gensim based models which cannot use GPUs and are required for word expansion) and GPU(for BERT inference).

All code and results are available at the following link. <https://github.com/Musarratpcr/CausalityDetection>.

In the following sub-sections, we shall provide the results obtained from various experiments in Stage 2 (threshold selection) and 3 (evaluation) of the setup(as shown in Figure 7.8).

7.2.2 Stage 2 - Threshold Selection Results

Following the process of preprocessing in Section 5.1, candidate triple extraction in Section 5.3.1, and causal candidate classification in Section 5.3.2, we calculated the cosine distance between the

candidate triples of the SemEval 2010 Task 8 test dataset against the six BERT models. Then using the truth values (labels) of each candidate triple from the test dataset, and the similarity score pertaining to the cosine distance, we individually evaluated the six BERT models, producing charts shown in Figure 7.10. We evaluated each threshold point, by connecting it with the inverse diagonal of the graph (From Precision=1 and recall=0 to Precision=0 and recall=1). We then calculated the area under this newly formed curve, and found out the threshold where this area was maximized. The average threshold value α then comes to 0.88, however, utilizing this average value in the multi-model CTTM would greatly affect the performance, by misclassifying instances for five individual models (more phrases will be classified as causal by bert-base-nli-mean-tokens, bert-base-nli-cls-token, bert-large-nli-cls-token, and less for bert-base-nli-max-tokens, and bert-large-nli-max-tokens). Instead, in the CTTM, we utilized the individual threshold values of each BERT model α_i , to classify instances, when compared to the corresponding embedding vector list.

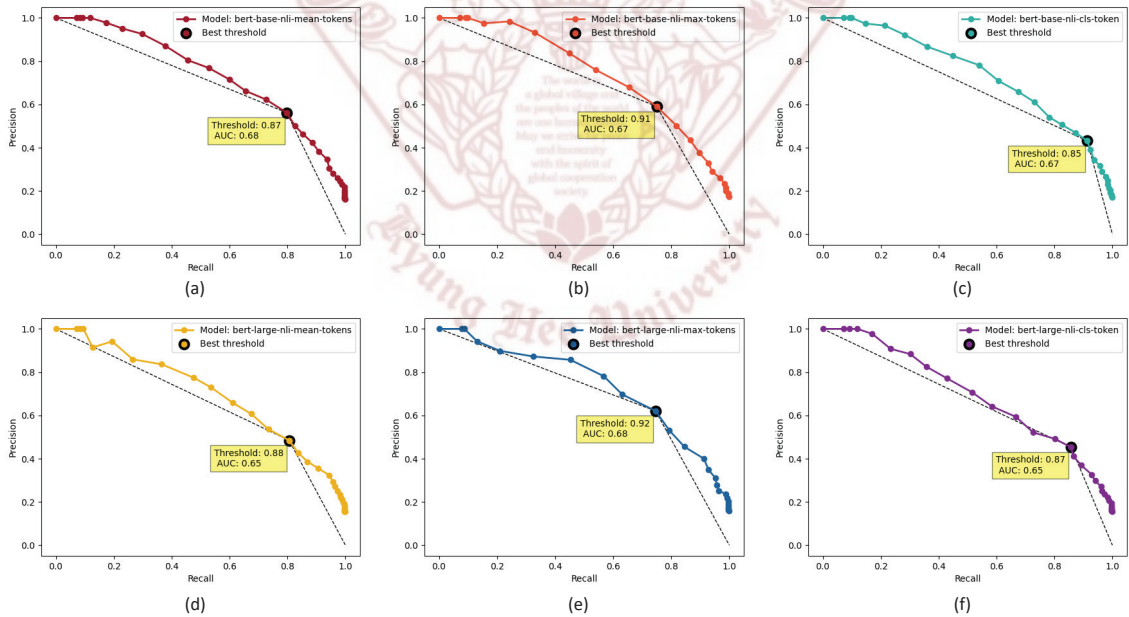


Figure 7.10: Precision recall curve for threshold selection (a) bert-base-nli-mean-tokens (b) bert-base-nli-max-tokens (c) bert-base-nli-cls-tokens (d) bert-large-nli-mean-tokens (e) bert-large-nli-max-tokens (f) bert-large-nli-cls-tokens.

In order to evaluate the performance of our generated triples and the selected threshold we then performed single-model, multi-model, and multi-model with feedback loop evaluation of three, as

yet, unseen datasets, the Asian Bayesian Network dataset and the two partitions for risk factors of Alzheimer's disease (AD1 and AD2). These are discussed as follows:

Some initial experiments, including evaluation of only verb expansion, and embedding vector generation using Word2Vec, comparison of six pre-trained BERT models (base-nli-mean-tokens, large-nli-mean-tokens, base-nli-max-tokens, large-nli-max-tokens, base-nli-cls-tokens, large-nli-cls-tokens), and application of BioBert embeddings [98] are explained with some detail in the following sections.

7.2.3 Experimental result with Word2Vec embeddings

Here, we performed some initial experiments to test the applicability and performance of Word2Vec based embedding vector generation process, for causal verbs and causal triples, in both training and test datasets. A summary of the results are shown in Table 7.2.

In Experiment 1, we extracted the causal verbs using the stanford POS tagger, from our training dataset. Without any expansion, we then applied word embedding on the causal verb, which was used to look up similar verbs in the SemEval test data set. In this iteration, we predicted 1318 sentences to be positively causal and 1399 sentences to be non-causal. From the predicted positive sentences, actual causal sentences were 205, and incorrect ones were 1113. The accuracy of this approach is 54.50% and recall 62.5%. However, the precision of this scenario is only 15.55% and F1 is 24.81%.

In Experiment 2, we expanded the causal verbs extracted in experiment 1 using Google News pre-trained model. Using word embedding, we transformed the extracted as well as the expanded causal verbs into word vectors. In the SemEval test data, using cosine similarity, 1453 sentences were classified as causal, with 210 correctly classified and 1243 incorrectly. After causal verb expansion the accuracy was dropped to 49.90%, precision to 14.45%, F1 to 23.58% but recall increased slightly to 64.40%. This indicates that word expansion from Google News pre-trained model has a very small impact on the classification process.

In Experiment 3, we switched the word expansion model to ConceptNet, with numberbatch embeddings, which provides semantically similar terms. In this iteration, we predicted 929 sentences to be causal and 1788 sentences to be non-causal. However, only 59 causal sentences were

Table 7.2: Initial Experiments with Word2Vec based embedding vector generation on SemEval Test dataset

Experiment	TP	FN	FP	TN	A(%)	P(%)	R(%)	F1(%)
1	205	123	1113	1276	54.50	15.55	62.5	24.81
2	210	118	1243	1146	49.90	14.45	64.02	23.58
3	59	269	870	1519	58.07	06.35	17.98	09.39

correctly predicted, with an accuracy of 58.07%, recall of 17.98% and lowest precision of 6.35% and F1 of 09.39% amongst all experiments. Causal terms are highly discriminable, while the words expanded with ConceptNet have higher diversity and lacks discrimination, which leads to the drastic decrease in the model performance [99]. The results obtained thus far have proved the in-applicability of Word2Vec based embedding vectors generation. The Word2Vec considered a word without its context and neighbor terms, which may lead to inappropriate vector generation. Therefore, we generated the embedding vectors via BERT models in the upcoming experiments.

7.2.4 Experimental Result with BERT Embeddings

In Experiment 7.2.4, like in the experiment 7.2.3 only verb was expanded. However, in this experiment the embedding vectors were generated using 6 BERT models to utilize sentence level embedding vector generation for a more contextual comparison. We compared 6 different BERT pre-trained models in terms of their performance on our test data set, with summary results shown in Table 7.3 [87,95]. The 6 BERT models (nli-base-mean-tokens, nli-large-mean-tokens, nli-base-max-tokens, nli-large-max-tokens, nli-base-cls-token, and nli-large-cls-token) differ in terms of their model size(base or large) and the pooling layer used at the end of their deep neural network(mean pooling word tokens, max pooling word tokens, or cls pooling sentence token). Experiment 4 pertains to the base form of the BERT model that uses mean token pooling, while Experiment 5 uses the large form of similar layered model. Likewise, Experiment 6 is the base model, while Experiment 7 is the large model, with max pooling layer. Finally, Experiment 8, and 9 are base and large models, respectively, with cls pooling layer. The result obtains in each experiments is shown in Table 7.3.

The result of these experiments show much improved performance, with experiment 4

(base model with mean pooling) showing the best accuracy(88.55%), precision(52.27%) and F1 (55.76%). The best recall(69.82%), is however, produced by the experiment 7 (large model with max pooling). On close inspection, we found experiment 7 to have correctly classified 229 sentences out of which 196 sentences were exactly similar to the True Positive results in experiment 4. However, the precision of experiment 7 is relatively small, due to the large number of False Positives.

Table 7.3: Setting 2 with BERT based embedding vector generation on SemEval Test dataset

Experiment	TP	FN	FP	TN	A(%)	P(%)	R(%)	F1(%)
4 - BERT-base-nli-mean-tokens	196	132	179	2210	88.55	52.27	59.76	55.76
5 - BERT-large-nli-mean-tokens	211	117	300	2089	84.65	41.29	64.33	50.30
6 - BERT-base-nli-max-tokens	227	101	633	1756	72.98	26.40	69.21	38.22
7 - BERT-large-nli-max-tokens	229	99	564	1825	75.60	28.88	69.82	40.86
8 - BERT-base-nli-cls-token	202	126	217	2172	87.38	48.21	61.59	54.08
9 - BERT-large-nli-cls-token	206	122	264	2125	85.79	43.83	62.80	51.63

Beyond these tests, it is also imperative that the generated embedding are tested on other text corpora for determining their ability to maintain acceptable performance, generally. Asia Bayesian Network and risk factors of Alzheimer's disease (AD) dataset were used to test this generalization. The results for the former are shown in Table. 7.4 and later in Table. 7.5. As shown in Table. 7.5 accuracy of each model decreases on Asia Bayesian Network as well as AD datasets. However, precision as well as recall of models shows a slight improvement on diverse datasets. In results for Experiment A.2 on the Asia Bayesian Network dataset, BERT nli-base-mean-tokens and BERT nli-large-mean-tokens show a precision of 100%, which is because of 0 false positives, however, this result is biased due to the very small number of identified causal triples.

These results paint an abysmal picture of the Experiment A.2 process. This is due to the fact that the verbs identified as causal through extraction from SemEval training dataset and their expansion are not able to capture all the causal sentences. These result partially support our novel methodology of incorporating the nominals (nouns and noun phrases) in the text producing the embedded vectors, thereby switching to causal quads for causal sentence identification. The intuition behind this arrangement, stems from the fact that causal sentences, implicitly contain semantic relationships between the cause and effect entities. Addition of these entities in the causal rela-

tionship identification process would spread a wider net for causal sentence identification. This intuition has been materialized and empirically tested in the manuscript.

Table 7.4: Application of trained embedding on Asia Bayesian Network dataset

Scenario	TP	FN	FP	TN	A(%)	P(%)	R(%)	F1(%)
BERT nli-base-mean-tokens	2	45	0	38	47.06	100.00	4.26	08.16
BERT nli-large-mean-tokens	4	43	0	38	49.41	100.00	8.51	15.69
BERT nli-base-max-tokens	11	36	11	27	44.71	50.00	23.40	31.88
BERT nli-large-max-tokens	31	16	18	20	60.00	63.27	65.96	64.58
BERT nli-base-cls-token	6	41	1	37	50.59	85.71	12.77	22.22
BERT nli-large-cls-token	1	46	2	36	43.53	33.33	2.13	04.00

Table 7.5: Application of trained embedding on Risk Factors of Alzheimer’s Disease dataset

Scenario	TP	FN	FP	TN	A(%)	P(%)	R(%)	F1(%)
BERT nli-base-mean-tokens	53	423	16	316	45.67	76.81	11.13	19.45
BERT nli-large-mean-tokens	162	314	83	249	50.87	66.12	34.03	44.94
BERT nli-base-max-tokens	276	200	148	184	56.93	65.09	57.98	61.33
BERT nli-large-max-tokens	282	194	194	138	51.98	59.24	59.24	59.24
BERT nli-base-cls-token	110	366	50	282	48.51	68.75	23.11	34.59
BERT nli-large-cls-token	176	300	84	248	52.48	67.69	36.97	47.83

7.2.5 Experimental Result with BioBERT Embeddings

The experiments performed in Appendix 7.2.4 are repeated by replace the BERT model with BioBert for generated trigger and candidate embeddings for comparing their similarities. As mentioned earlier, the trigger in the form of triple {noun, verb, noun} was extracted from SemEval training datasets, and the verb terms were expanded with Google news model to extended the converge of the triggers. We calculate precision recall curve as shown in 7.11, to identify the similarity cut off value of 0.96 for classifying a triple as causal and non-casual. However, the performance of the BioBert Embeddings are very low on the test dataset as shown in Table 7.6. The unexpected performance of the BioBERT based embeddings is mainly due to the fact our test dataset contains non-clinical concepts along with clinical concepts. Therefore, we used Bert models instead of BioBert for our experiments and evaluations.

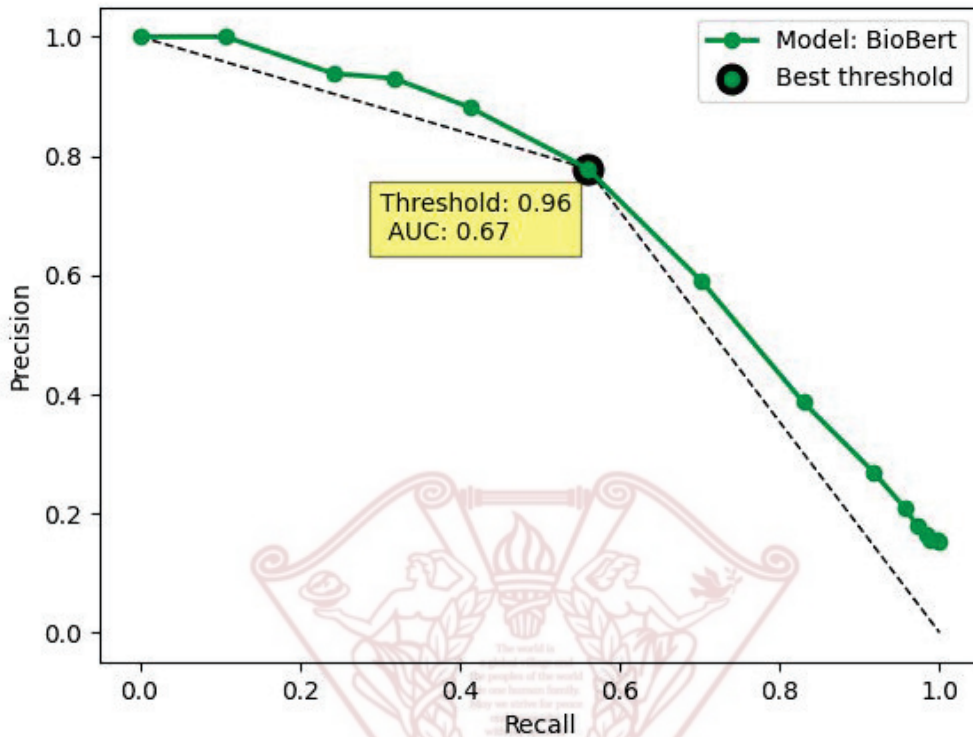


Figure 7.11: Precision recall curve for threshold selection for BioBERT.

Table 7.6: Application of BioBERT Embedding on Test Datasets

Dataset	TP	FN	FP	TN	A(%)	P(%)	R(%)	F1(%)
Asia	5	42	6	33	44.19	45.45	10.64	17.24
AD	29	494	10	331	41.67	74.36	05.54	10.32

The rest of the experimental setup can be categorized into 3 stages, as shown in Figure 7.8, where each following stage, receives data from all previous stages.

7.2.6 Single Model Evaluation

In the Asia Bayesian Network dataset from a total of 86 qualifying triples, 47 are actual causal(54.65%) and 38 are non-causal(44.18%). The results achieved by each BERT model on this dataset are shown in Table 7.7. BERT models, utilizing the complete phrase as a token and

then cls for pooling at the final layer, show good performance, when compared with the others. Overall, the best values for accuracy, precision, recall and F1 are achieved by these models, however, the BERT nli-large-mean-tokens closely follows the classification performance. However, the results for base models with mean tokens and max tokens, indicate very bad performance with F1 measure under 28% (caused by the low performance of recall obtained by these models).

Table 7.7: Application of trained embedding on Asia Bayesian Network dataset

Legend: TP is True positive, FN is False Negative, FP is False Positive, TN is True Negative, A is accuracy, P is precision, R is recall, and F1 is F1 Score

Scenario	TP	FN	FP	TN	A(%)	P(%)	R(%)	F1(%)
BERT nli-base-mean-tokens	8	39	5	34	48.84	61.54	17.02	26.67
BERT nli-large-mean-tokens	37	10	26	13	58.14	58.73	78.72	67.27
BERT nli-base-max-tokens	9	38	10	29	44.19	47.37	19.14	27.27
BERT nli-large-max-tokens	21	26	14	25	53.49	60.00	44.68	51.22
BERT nli-base-cls-token	34	13	19	20	62.79	64.15	72.34	68.00
BERT nli-large-cls-token	38	9	26	13	59.30	59.38	80.85	68.47

In absolute terms, the classification performance for the AD dataset in terms of accuracy, and F1 measure is lower than the Asia Bayesian Network dataset, at par for precision, and higher for recall. Comparison amongst the six models shows some similarity with the previous results. Causal classification of AD1 shown in Table 7.8, achieves better performance, in terms of its accuracy, recall, and F1 for the two cls-token models, with the large version achieving the best results. The performance of other models, lacks behind substantially with F1 rates between 34% and 49%. The precision rates of these six models, are however, within 5.15 percentage points, which indicates that the ability of each model to correctly identify the actual causal phrase, when a triple is classified as causal, is similarly good (or bad). Another important metric to analyze these results is to look at the recall rates, which in the case of Asia Bayesian Network dataset, were able to correctly identify 80.85% of the actual causal instances, however, for AD1 only identify 61.80%, in the best case. For the AD2 dataset, performance metrics shown in Table 7.9, indicate the best recall rate of 67.19%, which is better than the results for AD1 but substantially smaller than Asia Bayesian Network dataset. The best F1 rates for AD2 are achieved by the base version of the BERT cls token based model.

These results provide empirical proof for causality detection, based on causal phrase extrac-

Table 7.8: Application of trained embedding on Risk Factors of Alzheimer's Disease Split 1

Scenario	TP	FN	FP	TN	A(%)	P(%)	R(%)	F1(%)
BERT nli-base-mean-tokens	62	205	36	129	44.21	63.27	23.22	33.97
BERT nli-large-mean-tokens	111	156	80	85	45.37	58.12	41.57	48.47
BERT nli-base-max-tokens	72	195	45	120	44.44	61.54	26.97	37.50
BERT nli-large-max-tokens	80	187	54	111	44.21	59.70	29.96	39.90
BERT nli-base-cls-token	157	110	100	65	51.39	61.09	58.80	59.92
BERT nli-large-cls-token	165	102	104	61	52.31	61.34	61.80	61.57

tion and expansion. However, selection of a single model, based on these results alone, would not resolve the problem of causality detection in clinical text, where it is critical to identify most if not all actual causal sentences. Hence a deeper look at the coverage of these six BERT models, in terms of correctly classifying the actual causal instances is necessary.

Table 7.9: Application of trained embedding on Risk Factors of Alzheimer's Disease Split 2

Scenario	TP	FN	FP	TN	A(%)	P(%)	R(%)	F1(%)
BERT nli-base-mean-tokens	60	196	27	149	48.38	68.97	23.44	34.99
BERT nli-large-mean-tokens	128	128	70	106	54.17	64.65	50.00	56.39
BERT nli-base-max-tokens	74	182	37	139	49.31	66.67	28.91	40.33
BERT nli-large-max-tokens	88	168	54	122	48.61	61.97	34.38	44.22
BERT nli-base-cls-token	166	190	94	82	57.41	63.85	64.84	64.34
BERT nli-large-cls-token	172	84	111	65	54.86	60.78	67.19	63.82

The associations between the results achieved by six bert models on combined triple phrases from Asia Bayesian Network and Risk Factors of Alzheimer's Disease datasets is shown in Figure 7.12. Amongst the 950 candidate triples, 754 have been classified as causal by one or more of the BERT NLI models. The Base-Mean classifier, is unable to uniquely classify any candidate triple as causal, however, the other five models, classify 153 instances as causal. With classification intersection 2, 156 candidate triples are classified by a combination of only two models uniquely classify an instance as causal. Extending this calculation on the numbers achieved via UpSet analysis, unique coverage rate from degree 1-6 are 153(20.29%), 156(20.69%), 139 (18.44%), 127 (16.84%), 70 (9.28%), and 109 (14.46%), respectively. The actual causal triples in the 950 candidate triples are 570. True positive classification numbers for the six models with degree 1-6 are

83 (14.56%), 96 (16.84%), 86 (15.09%), 76 (13.33%), 45 (7.89%), and 70 (12.28%), respectively. These results indicate that single model application of anyone of the six BERT models, will have low candidate classification coverage and even lower true positive rates.

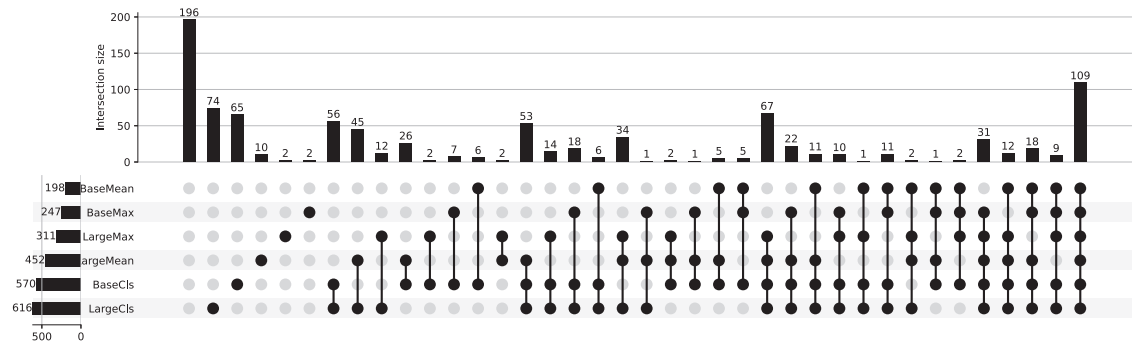


Figure 7.12: UpSet analysis of BERT model classification coverage for a combined list of Risk Factors of Alzheimer’s Disease and Asia Bayesian Network dataset.

7.2.7 Multi-model Evaluation

Theoretical analysis of the results shown in Figure 7.12, indicate that when degree ≥ 6 , 109 phrases have been classified as causal, out of which 70 are actual causal. The accuracy of this classification is 43.26% and F-1 rate is 20.62%. For degree ≥ 5 , 179 phrases have been classified as causal, with 115 as true positive. The accuracy rate now, increases to 45.37%, while the F-1 goes up to 30.71%. Similarly, for degree ≥ 4 accuracy further increases to 48.00% and F-1 to 43.61%. For degree ≥ 3 , the accuracy becomes 51.47%, and F-1 54.58%. For degree ≥ 2 , accuracy further improves to 55.26%, and F-1 to 63.71%. Finally for degree ≥ 1 , at least 1 model classified 754 instances as causal, out of which 456 are actual causal. The accuracy increases to 56.63%, and F-1 to 68.88%. Matching the intuition, presented earlier, this analysis, also shows, that if atleast one model classifies a candidate phrase as causal, it should be accepted, to achieve the highest realistic performance.

Practical application of the multi-model methodology, where a phrase is considered causal, if atleast one model classifies it as such, produces the same result, showing an accuracy rate of 56.63% and F-1 as 68.88% for the combined dataset. Separately, the results for Asia Bayesian Net dataset show small improvement in their F-1 score (multi-model selected additional 3 correct

Table 7.10: Application of Multimodel Embedding on Test Datasets

Dataset	TP	FN	FP	TN	A(%)	P(%)	R(%)	F1(%)
AD1	210	57	132	33	56.25	61.40	78.65	68.97
AD2	205	51	138	38	56.25	59.78	80.08	68.45
Asia	41	6	28	11	60.47	59.42	87.23	70.69

causal phrases than the best results for BERT nli-large-cls-token on this dataset) and a slight drop in its accuracy, due to an increase in the number of True negatives (skewing the accuracy measure, towards positive results). Both AD1 and AD2 dataset, show substantial improvement of causality classification, with the application of multi-model technique. The number of correctly classified causal phrases in AD 1 have increased from 165 in the best case to 210, while for the AD2 have increase from 172 in the best single model application to 205 here. Overall the performance of multi-model classification on this dataset has brought it at par with the result of the other dataset. The F-1 measures for both AD1 and AD2 have increased, showing the correctness of the multi-model strategy for causality detection. However, the large number of false positives and true negatives, still leave a room for improvement of this model, which we resolved by additionally employing the feedback loop. The results for this upgrade are shared in the following evaluation.

7.2.8 The Feedback Loop Evaluation

The results achieved via the multi-model methodology presented in Section 7.2.7 are further enhanced through multiple active learning iterations. An iteration represents an execution (triples classification by applying the proposed causality mining methodology, including expert feedback) for an unseen dataset. As we have three test datasets (AD1, AD2, and Asia), therefore, we performed three active learning iterations.

In Iteration-1, the CTTM trained on the SemEval 2010 Task 8 training dataset was tested using the AD1 dataset. The six models in CTTM were updated by adding embedded vectors for the 314 causal triples verified by the expert and, removal of triples with similarity score α_i for the 28 marked as incorrectly classified. In the base version of the nli-mean-tokens model, 60 similar triples were removed, while in large version 190 triples were removed. Similarly, for the base and large version of the nli-max-tokens 94 and 143 triples were removed, respectively. Finally for

the cls-token version, 676 triples were removed from base and 626 triples from the large version. As shown in the Table 7.11, the accuracy of multi-model CTTM application on the AD2 dataset, shows minor improvement, in accuracy (from 56.25% to 60.87%), precision (from 59.78% to 60.43%), recall(80.08% to 98.44%), and F1 (68.45% to 74.86%), on incorporation of results from AD1.

In Iteration-2, the expert verified 368 classified causal triples as correct, while 49 were marked as non-causal. Based on this new set of causal triples, we again updated the CTTM before iteration 3, to further add the 368 embedding vectors and removed 175 triples from base version of the nli-mean-tokens, and 308 from the large version. For the nli-max-tokens 251 were removed from base version and 477 from large version. Finally, in the case of cls-token 804 were removed from base and 774 from large.

In Iteration-3, the evolved CTTM was applied on the Asia Bayesian Network Dataset, which registered small improvements on the multi-model results. Since this dataset is the smallest of the three, CTTM model evolution has very little impact on it. Addition of 759 triples in the original 1,246,975 embedded vectors from CTTM model before iteration 1, and removal of various others (between the minimum total of 235 triples removal from base nli-mean-tokens in 2 iterations and maximum of 1480 from base cls-token), increased the true positive from 38 in best case single model to 41 in multi-model, and finally to 42 in the third iteration.

Table 7.11: Feedback loop results on test datasets

Iteration	Dataset	Dataset Evaluation				Expert Evaluation	
		A	P	R	F1	Added to Embeddings	Added to Block List
1	AD1	56.25%	61.40%	78.65%	68.97%	314	28
2	AD2	60.88% (↑ 4.63)	60.43% (↑ 0.65)	98.44% (↑ 18.36)	74.89% (↑ 6.44)	268	49
3	Asia	61.63% (↑ 1.16)	60.00% (↑ 0.58)	89.36% (↑ 2.13)	71.79% (↑ 1.1)	58	12

A: Accuracy, P: Precision, R: Recall, The values in parenthesis represent rate of change from multi-model results.

7.2.9 Comparison with existing studies

In order to compare our methodology with an existing study, we utilized the methodology presented by [65] to classify sentences as causal or non-causal, from the AD1, AD2, and Asia dataset. However, since our methodology incorporates the datasets into the CTTM, using feedback loop, and because we want to maintain the unseen nature of these, so as not to contaminate the results,

we compared our iteration 1 result for AD1, iteration 2 result for AD2, and iteration 3 result for Asia dataset. At these specific points, the datasets are unseen and true test sets. The results for causal classification on the test datasets for both methodologies (Ning’s and proposed) are shown in Table 7.12. We observed that our implementation of Ning’s methodology [65], classifies all triples as causal achieving a recall rate of 100%. However, the accuracy, precision, and F_1 scores are decreasing by comparatively large margins. Hence, it is safe to conclude that even when starting with a well-identified set of causal verbs, word embedding by itself is not sufficiently able to evolve the causality classification model. On the other hand, our methodology is able to improve upon its results across iterations.

Table 7.12: Result comparison with Ning’s method on test datasets

Dataset	Ning’s Method Evaluation				Proposed Method Evaluation			
	A(%)	P(%)	R(%)	F1(%)	A(%)	P(%)	R(%)	F1(%)
AD1	61.81	61.81	100	76.39	56.25	61.40	78.65	68.97
AD2	59.26	59.26	100	74.42	60.88	60.43	98.44	74.89
Asia	54.65	54.65	100	70.68	61.63	60.00	89.36	71.79

7.2.10 Discussion

The main aim of this study is to develop a framework that can identify causal sentences in clinical text. The success criteria of this framework are dependent on correctly identifying most causal relationships, with some leeway available in incorrect classification of non-causal sentences as causal. Precision, recall, and their association in the form of F1 provides the metric to evaluate our proposed framework, in parts, as a whole, and with existing work. Application of this classification methodology can then enable an expert from the domain of healthcare and wellness, to be able to contextually summarize the contents of the clinical text. To this end, we extract the causal phrases from the causal sentences, which are larger in numbers but smaller in their participating linguistic elements (including two NP and one VP).

The results presented in section 7.2, provide the performance metrics for various steps leading up to our proposed multi-model classification with a feedback loop. In particular, the evaluation metrics for the Asian Bayesian Network, and two partitions of the Alzheimer’s disease datasets,

generally saw an increase, when moving from single model to multi-model and then to multi-model with a feedback loop.

The rationale for moving from using a single BERT NLI model to a multi-model application was established using UpSet analysis, presented in section 7.2.6. Additionally, the rationale for moving from multi-model to multi-model with a feedback loop can be naively established from intuition, however, it is far more beneficial to analyze the phrases which were originally classified by the machine learning models and then removed by the expert. As an example one of the triples identified by the multi-modal methodology from the AD1 dataset is “cancer = alcohol”. The origin of this triple can be traced back to the following instance:

“After adjusting for various socioeconomic and health variables, no significant differences were observed between hazardous drinking and type of cancer [PR = 0.99 = 0.83-1.17) in people with alcohol-related cancers compared to non-alcohol related cancers] and time since diagnosis [PR = 1.01 in people with a cancer diagnosed >5 years ago compared to those diagnosed ≤5 years ago].”

Stanford POS tagger (version 3.9.2) had incorrectly identified the symbol “=” as a feasible VP and since this fell between the two entities (cancer and alcohol), this triple was considered valid. The expert verified this triple as incorrect since it does not provide enough information to classify the original sentence as causal or non-causal. Hence the embedding for this triple and all others similar to it, with the threshold equal to or above, for each model, were then removed. This removal process is not dependent only on the VP, as in iteration 2, we observed additional triples with invalid VPs, such as “stroke = diabetes” and “alcohol [depression”. In iteration 3, none of the triples had a symbol as a VP.

Another triple identified by the expert as incorrect was “smoke monitored hypertension”. This triple contains a valid VP, tagged by the POS tagger as “VBN”. The original instance from which this triple was extracted is as follows

“The earlier advice to physicians still seems prudent and is briefly stated: 1) Try to avoid prescribing oral contraceptives for women over 35 years of age; 2) Women who smoke cigarettes should avoid using oral contraceptives, and users should not smoke; 3) Prescribe the formulation with the lowest dose and/or potency of estrogen that is effective and that does not cause unac-

ceptable "breakthrough" bleeding; 4) Women with hypertension should be carefully monitored, and women who develop hypertension while on oral contraceptives should be switched to another form of contraception, if possible."

In hindsight, intuitively, it is evident from the original text that the phrases "smoke", "monitored", and "hypertension" all belong to different contexts. However, the machine learning models are agnostic to such contexts, unless they can incorporate a very large number of sentence and document structuring rules. While there are other triples extracted from this instance, which may qualify the instance as causal or non-causal, it does not help fulfill our aim of identifying individual causal sentences from the classification of causal triples. Hence, we update the model, to only hold those triples which can represent causal relationships, from a wide variety of datasets. In iteration 3, only 12 triples were identified as incorrect by the expert, including "bronchitis smoke smoking" and "lung cancer secondhand smoking" (while lung cancer can be caused by secondhand smoking, this triple is missing the causal verb). Here it is pertinent to mention that by removing these triples, we are not changing our results but rather evolving the model for subsequent classification in unseen datasets. In the absence of active learning, our model would not be able to update itself and hence provide relatively mediocre results as discussed in section 7.2.7.

In iterations 2 and 3, incorrect triple embeddings similar to the ones identified in previous iterations are not included. The similarity is determined by converting the incorrect triples into embedding vectors and using the 6 BERT models to determine all embeddings which have cosine similarity above their respective thresholds. The correctly identified causal triples are added into the CTTM by appending their embeddings at the end. Additionally, on subsequent classifications, the data instances (a sentence, text excerpt, or a document) are classified using the evolved CTTM. This is why even after removing related embeddings the results obtained by including active learning are gradually increasing, even on unseen and minimally related datasets (AD2 in iteration 2 to Asia in iteration 3).

On a related note, the evaluation of our results has been performed using the labels of the test data, while the expert-provided feedback was used only to update the model. As a result some phrases such as "cancer associated alcohol", and "cancer rising alcohol", were classified by the machine learning model and the expert as causal, however, the dataset had the associated sentence

labeled as non-causal. Since CTTM is direction agnostic it is unable to distinguish between various forms of the causal phrases such as “cause triggers effect” and “effect triggered by cause”. As an example the triple “cancer associated alcohol” has been extracted from the following instance:

“The results showed that frequent intake of fruits, chicken, fish and alcohol drinking were associated with risk for colorectal cancer.” Here, the triple has been correctly identified, since one of the verbs between “cancer” and “alcohol” is a verb token “associated” which generate a triple <cancer, associated, alcohol>. During triples classification step, the triple was also identified as causal by the CTTM and the human expert, however, the source dataset marks it as non-causal. Thus while the triple itself is causal, the originating instance is non-causal (the dataset labels it as “cancer x alcohol”), which negatively affects the evaluations and reduces the performance of the proposed solution.

The causality classification methodology presented in this manuscript attempts to alleviate problems caused by discrepancies in causally valid POS tagging, triple expansion (which can include non-causal triples), and other operations. Through the use of active learning, we have observed an increase in the performance of our proposed methodology while reduce negative detected causal triple. Although, we have reduced the expert’s involvement in the causality classification process, substantially, when compared with the previous studies, further reduction is possible through the use of specialized POS taggers, contextual triple expansions, better sentence embedding generation, and similarity measures.

7.3 Rules Generation Results

7.3.1 Experimental Setup

The rule generation methodology presented in Chapter 6 represents theoretical steps for concepts values and operator extraction. The realization of the proposed module is achieved by processing the Hypertension guideline [78] for concepts and associated value extraction. The guidelines consist of 87 recommendation sentences among 278 total sentences. A human annotator annotated each recommendation sentence of the guideline for concept, operator, and its associated value. A total of 71 concepts were annotated with associated values. These annotations are used as ground

truth to be compared with values automatically extracted by the proposed methodology for rule generation.

7.3.2 Concepts Values Extraction Results

The proposed technique for concepts' value extraction is evaluated on annotated Hypertension guideline [78]. Among the annotated 71 concepts, the proposed module correctly identified values for 65 concepts, while missed out 6 values resulted in 91.55% accuracy for value extraction as shown in Table 7.13.

Table 7.13: Concepts value extraction result

Total Concepts	Identified Values	Missed Values	Accuracy
71	65	6	91.55%

The proposed solution accurately identified concepts value in both cases where a valued is preceded or followed by the target concept. Also, it efficiently detected the range of values for a concept. For example in a sentence "There is strong evidence to support treating hypertensive persons aged 60 years or older to a BP goal of less than 150/90 mm Hg and hypertensive persons 30 through 59 years of age to a diastolic goal of less than 90 mm Hg ; however , there is insufficient evidence in hypertensive persons younger than 60 years for a systolic goal , or in those younger than 30 years for a diastolic goal , so the panel recommends a BP of less than 140/90 mm Hg for those groups based on expert opinion." the system accurately detected age = 30-59 among other concepts and their values.

By analyzing the concepts where our proposed solution missed their value identification, we found out that in all cases the value is located far away from the concept in the source sentence. For example in the sentence "Although treatment with an ACEI or ARB may be beneficial in those older than 75 years , use of a thiazide-type diuretic or CCB is also an option for individuals with CKD in this age group." the concept of interest "age" is located at the end of the sentence while its value is at the start of the sentence. Our solution mainly focuses on concept neighbor tokens for value and operator identification to maintain context of the targeted concepts.

The identified concepts, their operators and values are used for completing the condition as

well as the action part of the rule as final knowledge. The acquired transparent knowledge can be used by automated clinical systems such as CDSS for clinical decision support or can be utilized by human experts for better decision making and quality improvements.

7.4 End-to-end Evaluation

The end-to-end evaluation of the proposed methodology is performed on three different guidelines from the diabetes domain [100–102]. Guideline one is developed by American Diabetes Association (ADA), Guideline two by National Institute for Health and Care Excellence (NICE), and guideline three by the Scottish Intercollegiate Guideline Network (SIGN). The aim is to check the effectiveness of the methodology on the diverse nature of the clinical text. The experimental setup and results achieved are shown in Table 7.14.

Table 7.14: End-to-end methodology results

Process	Guideline1 [100]	Guideline2 [101]	Guideline3 [102]	Total
Sentences	367	1805	279	2451
Extracted Triples	1731	1142	10226	13099
Unique Triples	1602	948	8872	11422
Medical Triples	541	320	7765	3215
Extracted Rules	29	7	13	49

The evaluation of the extracted knowledge (Production rules) is performed on a de-identified real patient dataset from one of our collaborative hospital ². The dataset consists of total 302 patient instances with four possible labels (Non Diabetes Mellitus, Recheck, Pre-Diabetes Mellitus, and Diabetes Mellitus) and 12 features, including current and previous values of FPG, HbA1c, OGTT, PPG, Sign and Symptoms, and RPG values. The classes of the instances are distributed with 63:49:60:130 ratios.

The classification results achieved by the extracted knowledge compared to expert decision as ground truth is shown in Figure 7.13. As we can see, the extracted rules achieved overall accuracy of 71.79% compared to expert decisions. However the analysis of the result revealed that 16.83% time expert provide final decision as “Recheck” while there is no rule in the guidelines related to

²<https://www.cmcseoul.or.kr/en.common.main.main.sp>

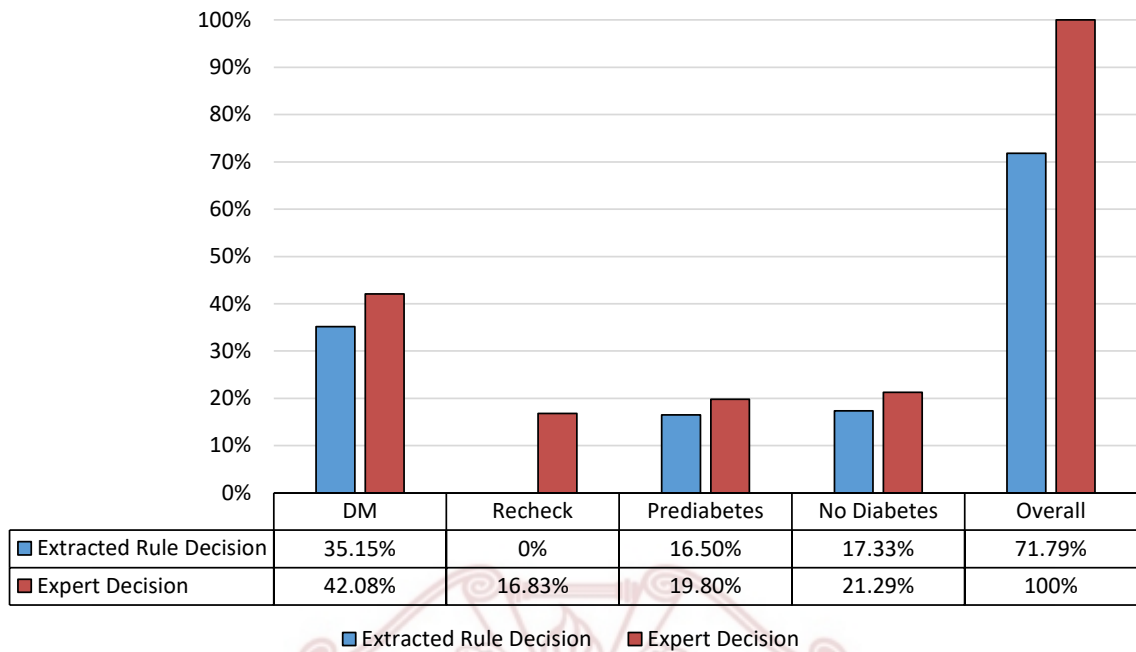


Figure 7.13: End-to-end methodology evaluation.

this class. All three guidelines assign one the three possible classes. Thus we can conclude that our methodology performed reasonably well and it can be further improved by multiple active learning iterations.

8.1 Conclusion

The drastic increase in healthcare data availability, advancement of artificial intelligence, and computing technologies can make AI-integrated healthcare systems possible. The intelligence of the AI systems mainly relies on the training data, while the majority of the data in the healthcare domain are stored in unstructured format due to the ease of use of the data. The unstructured format of the data makes it difficult for an automatic system to process, understand, and extract decision-making logic out of the data. This necessitates an end-to-end methodology of knowledge extraction from unstructured clinical data. Existing natural language processing solutions mainly focused on a single aspect such as text classification, entity extraction, or relation extraction of the knowledge acquisition process flow. Therefore, this dissertation proposed an end-to-end methodology for extracting machine-readable and transparent knowledge from unstructured clinical data. The methodology process the input documents in three sequential steps. First, the document text is classified into two main categories, recommendation and non-recommendation sentences, based on the importance of content presented by each sentence. The recommendation sentences are further processed for clinical entity and their relation extraction. Second, the clinical concepts with a cause-effect relationship are extracted. Finally, the causal concepts are processed for knowledge acquisition in the production rule format.

Clinical text classification is one of the widely explored research areas. However, there exists a huge gap between research work and real field applications. Clinical text classification researchers are mainly inclined towards advanced AI methods such as deep learning while real field applications still preferred pattern bases approaches due to their ease of use and decision transparency [32]. In this dissertation, we enhanced the pattern-based approaches by incorpo-

rating artificial intelligence and machine assistance in the manual as well as automatic pattern extractions process. Also, we mitigate the lack of generalization issue of the extracted patterns by presenting POS and UMLS-based patterns. The resultant patterns increased the classification performance compared to traditional as well as advanced machine learning-based approaches.

The recommendation tagged sentences from the text classification step are further processed for clinical concepts and their relation extraction. Clinicians are mainly interested to find the causes and effects of various clinical procedures, therefore, this dissertation explored the cause-effect relationship of the concepts. We proposed a novel methodology by leveraging the applications of transfer learning and active learning methodologies. The recommendation sentences are evaluated for clinical concepts and transformed into triples of the form $\langle \text{Noun phrase, Verb Phrase, Noun Phrase} \rangle$ where the noun phrase indicates the clinical terms and the verb phrase represents the causal trigger. The extracted triple phrases are transformed into embedding vectors using a set of pre-trained BERT models. The generated embedding is matched with causal triple embeddings to find the similarity of the triple with causal triples using the multi-model approach. A candidate triples having maximum similarity higher than the threshold value are classified as causal triples. By incorporating active learning methodology, the casual classified triples are verified by a human expert and feedback to the causal triple embeddings to enhance the model accuracy for the subsequent runs.

We further processed the causal triples for rules generation. A causal triple can either represents $\langle \text{cause, trigger, effect} \rangle$ or $\langle \text{effect, trigger, cause} \rangle$. We mapped the causal concept of the triple to the condition part while the effect concept to the conclusion part of the production rule. However, the condition part of the production rule contains the condition key, operation, and its value. Therefore, we evaluated the source senescent of the causal triple for possible operator and value extraction. Triple rules with the same source sentence are combined into a single rule to get more comprehensive knowledge.

The presented research is designed to extract valuable knowledge from unstructured clinical resources. It assists clinical experts in coping with the surfeit of unstructured clinical text [103, 104]. The acquired knowledge can directly be used in clinical decision support systems for better clinical decisions and healthcare quality improvement [10, 104]. It can also be utilized by

human experts to increase their intuitions and decision-making capabilities. Additionally, individual solutions of the methodology can be used as a subpart of other applications, such as patient health summary generation from associated unstructured documents.

8.2 Future Direction

In future, the presented clinical knowledge extraction pipeline can be further enhance by replacing individual modules with other state-of-the-art methods. Such as the classification model can be extended to multi-class classification. Similarly, the causality mining module can be enhanced by applying other algorithms such as GPT. Also, other relationships such as “improves”, “reveals“, etc can be include which will increase the scope and coverage of the the resultant knowledge. Additionally, the proposed methodology produces production rules as final knowledge which is easy to use and is understandable to both machines as well as human beings. However, the acquired knowledge can be represented in more feature-rich models such as knowledge graphs so that we can better preserve the semantics of concepts and their relationships. The knowledge graph will make us able to automatically enlarge and enhance the underlying knowledge through knowledge graph completion methodologies.

Bibliography

- [1] B. Meskó and M. Görög, “A short guide for medical professionals in the era of artificial intelligence,” *NPJ digital medicine*, vol. 3, no. 1, pp. 1–8, 2020.
- [2] S. Bahri, N. Zoghلامي, M. Abed, and J. M. R. Tavares, “Big data for healthcare: a survey,” *IEEE access*, vol. 7, pp. 7397–7408, 2018.
- [3] M. Assale, L. G. Dui, A. Cina, A. Seveso, and F. Cabitza, “The revival of the notes field: leveraging the unstructured content in electronic health records,” *Frontiers in medicine*, vol. 6, p. 66, 2019.
- [4] G. Saposnik, D. Redelmeier, C. C. Ruff, and P. N. Tobler, “Cognitive biases associated with medical decisions: a systematic review,” *BMC medical informatics and decision making*, vol. 16, no. 1, pp. 1–14, 2016.
- [5] A. Berger and L. L. Weed, “Opening the black box of clinical judgement.” *BMJ: British Medical Journal: International Edition*, vol. 319, no. 7220, pp. 1279–1279, 1999.
- [6] A. K. Lauer and D. A. Lauer, “The good doctor: more than medical knowledge & surgical skill,” *Annals of eye science*, vol. 2, 2017.
- [7] M.-H. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki, and D. K. Grunwell, “Health big data analytics: current perspectives, challenges and potential solutions,” *International Journal of Big Data Intelligence*, vol. 1, no. 1-2, pp. 114–126, 2014.
- [8] H. Hematialam and W. W. Zadrozny, “Identifying condition-action statements in medical guidelines: Three studies using machine learning and domain adaptation,” 2021.

- [9] H. Dalianis, *Clinical text mining: Secondary use of electronic patient records*. Springer Nature, 2018.
- [10] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn *et al.*, “Clinical information extraction applications: a literature review,” *Journal of biomedical informatics*, vol. 77, pp. 34–49, 2018.
- [11] P. M. Amisha, M. Pathania, and V. K. Rathaur, “Overview of artificial intelligence in medicine,” *Journal of family medicine and primary care*, vol. 8, no. 7, p. 2328, 2019.
- [12] R. Egger and E. Gokce, “Natural language processing (nlp): An introduction,” in *Applied Data Science in Tourism*. Springer, 2022, pp. 307–334.
- [13] P. Johri, S. K. Khatri, A. T. Al-Taani, M. Sabharwal, S. Suvanov, and A. Kumar, “Natural language processing: History, evolution, application, and future work,” in *Proceedings of 3rd International Conference on Computing Informatics and Networks*. Springer, 2021, pp. 365–375.
- [14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.
- [15] I. Spasic, G. Nenadic *et al.*, “Clinical text data in machine learning: systematic review,” *JMIR medical informatics*, vol. 8, no. 3, p. e17984, 2020.
- [16] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, V. Osmani *et al.*, “Natural language processing of clinical notes on chronic diseases: systematic review,” *JMIR medical informatics*, vol. 7, no. 2, p. e12239, 2019.
- [17] N. Chintalapudi, G. Battineni, M. Di Canio, G. G. Sagaro, and F. Amenta, “Text mining with sentiment analysis on seafarers’ medical documents,” *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100005, 2021.

- [18] D. A. Davis and A. Taylor-Vaisey, "Translating guidelines into practice: a systematic review of theoretic concepts, practical experience and research evidence in the adoption of clinical practice guidelines," *Cmaj*, vol. 157, no. 4, pp. 408–416, 1997.
- [19] J. Fox, V. Patkar, I. Chronakis, and R. Begent, "From practice guidelines to clinical decision support: closing the loop," *Journal of the Royal Society of Medicine*, vol. 102, no. 11, pp. 464–473, 2009.
- [20] P. Bose, S. Srinivasan, W. C. Sleeman, J. Palta, R. Kapoor, and P. Ghosh, "A survey on recent named entity recognition and relationship extraction techniques on clinical texts," *Applied Sciences*, vol. 11, no. 18, p. 8319, 2021.
- [21] M. Cui, R. Bai, Z. Lu, X. Li, U. Aickelin, and P. Ge, "Regular expression based medical text classification using constructive heuristic approach," *IEEE Access*, vol. 7, pp. 147 892–147 904, 2019.
- [22] Y. Zhou, C. Ju, J. H. Caufield, K. Shih, C. Chen, Y. Sun, K.-W. Chang, P. Ping, and W. Wang, "Clinical named entity recognition using contextualized token representations," *arXiv preprint arXiv:2106.12608*, 2021.
- [23] Z. Li, Z. Yang, C. Shen, J. Xu, Y. Zhang, and H. Xu, "Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–8, 2019.
- [24] P. Li and K. Mao, "Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts," *Expert Systems with Applications*, vol. 115, pp. 512–523, 2019.
- [25] M. A. Murtaugh, B. S. Gibson, D. Redd, and Q. Zeng-Treitler, "Regular expression-based learning to extract bodyweight values from clinical notes," *Journal of biomedical informatics*, vol. 54, pp. 186–190, 2015.
- [26] B. Jang, M. Kim, G. Harerimana, S.-u. Kang, and J. W. Kim, "Bi-lstm model to increase accuracy in text classification: combining word2vec cnn and attention mechanism," *Applied Sciences*, vol. 10, no. 17, p. 5841, 2020.

- [27] M. Thangaraj and M. Sivakami, "Text classification techniques: A literature review." *Interdisciplinary Journal of Information, Knowledge & Management*, vol. 13, 2018.
- [28] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.
- [29] M. Jiang, Y. Liang, X. Feng, X. Fan, Z. Pei, Y. Xue, and R. Guan, "Text classification based on deep belief network and softmax regression," *Neural Computing and Applications*, vol. 29, no. 1, pp. 61–70, 2018.
- [30] S. Xu, "Bayesian naïve bayes classifiers to text classification," *Journal of Information Science*, vol. 44, no. 1, pp. 48–59, 2018.
- [31] D. Cai, N. Garg, M. Dobrzynski, W.-Q. Guo, A. Khanna, and N. Xu, "Content pattern based automatic document classification," Jul. 14 2020, uS Patent 10,713,306.
- [32] S. Fu, D. Chen, H. He, S. Liu, S. Moon, K. J. Peterson, F. Shen, L. Wang, Y. Wang, A. Wen *et al.*, "Clinical concept extraction: a methodology review," *Journal of Biomedical Informatics*, p. 103526, 2020.
- [33] L. Yao, C. Mao, and Y. Luo, "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks," *BMC medical informatics and decision making*, vol. 19, no. 3, p. 71, 2019.
- [34] R. I. Doğan, A. Névél, and Z. Lu, "A context-blocks model for identifying clinical relationships in patient records," *BMC bioinformatics*, vol. 12, no. 3, pp. 1–11, 2011.
- [35] S. Zhao, T. Liu, S. Zhao, Y. Chen, and J.-Y. Nie, "Event causality extraction based on connectives analysis," *Neurocomputing*, vol. 173, pp. 1943–1950, 2016.
- [36] R. Girju, D. I. Moldovan *et al.*, "Text mining for causal relations." in *FLAIRS conference*, 2002, pp. 360–364.
- [37] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

- [38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [39] G. A. Miller, *WordNet: An electronic lexical database*. MIT press, 1998.
- [40] R. Rahman and C. K. Reddy, "Electronic health records: A survey," *Healthcare Data Analytics*, vol. 36, p. 21, 2015.
- [41] O. Müller, I. Junglas, S. Debortoli, and J. vom Brocke, "Using text analytics to derive customer service management benefits from unstructured data," *MIS Quarterly Executive*, vol. 15, no. 4, pp. 243–258, 2016.
- [42] B. Percha, "Modern clinical text mining: A guide and review," *Annual Review of Biomedical Data Science*, vol. 4, pp. 165–187, 2021.
- [43] K. S. Jones, "Natural language processing: a historical review," *Current issues in computational linguistics: in honour of Don Walker*, pp. 3–16, 1994.
- [44] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *arXiv preprint arXiv:1708.05148*, 2017.
- [45] K. Kaiser, A. Seyfang, and S. Miksch, "Identifying actions described in clinical practice guidelines using semantic relations," in *KR4HC 2010-2nd International Workshop on Knowledge Representation for Health Care*. Citeseer, 2010, pp. 99–108.
- [46] R. Wenzina and K. Kaiser, "Identifying condition-action sentences using a heuristic-based information extraction method," in *Process Support and Knowledge Representation in Health Care*. Springer, 2013, pp. 26–38.
- [47] H. Hematialam and W. Zadrozny, "Identifying condition-action statements in medical guidelines using domain-independent features," *arXiv preprint arXiv:1706.04206*, 2017.
- [48] S. Priyanta, S. Hartati, A. Harjoko, and R. Wardoyo, "Comparison of sentence subjectivity classification methods in indonesian news," *International Journal of Computer Science and Information Security*, vol. 14, no. 5, p. 407, 2016.

- [49] C. F. Meyer, *Introducing English Linguistics International Student Edition*. Cambridge University Press, 2010.
- [50] R. M. Kaplan and G. Berry-Rogghe, “Knowledge-based acquisition of causal relationships in text,” *Knowledge Acquisition*, vol. 3, no. 3, pp. 317–337, 1991.
- [51] K. Raja, S. Subramani, and J. Natarajan, “Ppinterfinder—a mining tool for extracting causal relations on human proteins from literature,” *Database*, vol. 2013, 2013.
- [52] R. Girju, “Automatic detection of causal relations for question answering,” in *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, 2003, pp. 76–83.
- [53] G. A. Miller, “Wordnet: A lexical database for english,” *Commun. ACM*, vol. 38, no. 11, p. 39–41, Nov. 1995. [Online]. Available: <https://doi.org/10.1145/219717.219748>
- [54] Q.-C. Bui, B. Ó. Nualláin, C. A. Boucher, and P. M. Sloot, “Extracting causal relations on hiv drug resistance from literature,” *BMC bioinformatics*, vol. 11, no. 1, pp. 1–11, 2010.
- [55] S. V. Cole, M. D. Royal, M. G. Valtorta, M. N. Huhns, and J. B. Bowles, “A lightweight tool for automatically extracting causal relationships from text,” in *Proceedings of the IEEE SoutheastCon 2006*. IEEE, 2006, pp. 125–129.
- [56] S. Doan, E. W. Yang, S. S. Tilak, P. W. Li, D. S. Zisook, and M. Torii, “Extracting health-related causality from twitter messages using natural language processing,” *BMC medical informatics and decision making*, vol. 19, no. 3, pp. 71–77, 2019.
- [57] S. Alashri, J.-Y. Tsai, A. R. Koppela, and H. Davulcu, “Snowball: extracting causal chains from climate change text corpora,” in *2018 1st International Conference on Data Intelligence and Security (ICDIS)*. IEEE, 2018, pp. 234–241.
- [58] J. Yang, S. C. Han, and J. Poon, “A survey on extraction of causal relations from natural language text,” *Knowledge and Information Systems*, pp. 1–26, 2022.

- [59] D.-S. Chang and K.-S. Choi, "Causal relation extraction using cue phrase and lexical pair probabilities," in *International Conference on Natural Language Processing*. Springer, 2004, pp. 61–70.
- [60] E. Blanco, N. Castell, and D. Moldovan, "Causal relation extraction," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.
- [61] N. Asghar, "Automatic extraction of causal relations from natural language texts: a comprehensive survey," *arXiv preprint arXiv:1605.07895*, 2016.
- [62] T. N. De Silva, X. Zhibo, Z. Rui, and M. Kezhi, "Causal relation identification using convolutional neural networks and knowledge based features," *International Journal of Computer and Systems Engineering*, vol. 11, no. 6, pp. 696–701, 2017.
- [63] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," in *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, 2014, pp. 2335–2344.
- [64] T. H. Nguyen and R. Grishman, "Relation extraction: Perspective from convolutional neural networks," in *Proceedings of the 1st workshop on vector space modeling for natural language processing*, 2015, pp. 39–48.
- [65] N. An, Y. Xiao, J. Yuan, Y. Jiaoyun, and G. Alterovitz, "Extracting causal relations from the literature with word vector mapping," *Computers in biology and medicine*, vol. 115, p. 103524, 11 2019.
- [66] D. Redd, J. Kuang, A. Mohanty, B. E. Bray, and Q. Zeng-Treitler, "Regular expression-based learning for mets value extraction," *AMIA Summits on Translational Science Proceedings*, vol. 2016, p. 213, 2016.
- [67] G. Zheng, S. Mukherjee, X. L. Dong, and F. Li, "Opentag: Open attribute value extraction from product profiles," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1049–1058.

- [68] T. Cai, L. Zhang, N. Yang, K. K. Kumamaru, F. J. Rybicki, T. Cai, and K. P. Liao, "Extraction of emr numerical data: an efficient and generalizable tool to extend clinical research," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–7, 2019.
- [69] M. Hussain, J. Hussain, T. Ali, S. I. Ali, H. S. M. Bilal, S. Lee, and T. Chung, "Text classification in clinical practice guidelines using machine-learning assisted pattern-based approach," *Applied Sciences*, vol. 11, no. 8, p. 3296, 2021.
- [70] M. Hussain, J. Hussain, T. Ali, and S. Lee, "An empirical method of automatic pattern extraction for clinical text classification," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 5292–5295.
- [71] S. K. Srivastava, S. K. Singh, and J. S. Suri, "State-of-the-art methods in healthcare text classification system: Ai paradigm," *Front Biosci*, vol. 25, pp. 646–672, 2020.
- [72] M. Hussain and S. Lee, "Information extraction from clinical practice guidelines: A step towards guidelines adherence," in *International Conference on Ubiquitous Information Management and Communication*. Springer, 2019, pp. 1029–1036.
- [73] W. Ali, W. Zuo, R. Ali, X. Zuo, and G. Rahman, "Causality mining in natural languages using machine and deep learning techniques: A survey," *Applied Sciences*, vol. 11, no. 21, p. 10064, 2021.
- [74] M. Hussain, F. A. Satti, J. Hussain, T. Ali, S. I. Ali, H. S. M. Bilal, G. H. Park, S. Lee, and T. Chung, "A practical approach towards causality mining in clinical text using active transfer learning," *Journal of Biomedical Informatics*, vol. 123, p. 103932, 2021.
- [75] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.
- [76] D. Cai, N. Garg, M. Dobrzynski, W.-Q. Guo, A. Khanna, and N. Xu, "Content pattern based automatic document classification," Mar. 28 2019, uS Patent App. 15/713,445.

- [77] D. D. A. Bui and Q. Zeng-Treitler, "Learning regular expressions for clinical text classification," *Journal of the American Medical Informatics Association*, vol. 21, no. 5, pp. 850–857, 2014.
- [78] P. A. James, S. Oparil, B. L. Carter, W. C. Cushman, C. Dennison-Himmelfarb, J. Handler, D. T. Lackland, M. L. LeFevre, T. D. MacKenzie, O. Ogedegbe *et al.*, "2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the eighth joint national committee (jnc 8)," *Jama*, vol. 311, no. 5, pp. 507–520, 2014.
- [79] B. Shekar and G. Dagnew, "Grid search-based hyperparameter tuning and classification of microarray cancer data," in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*. IEEE, 2019, pp. 1–8.
- [80] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.
- [81] A. R. Aronson and F.-M. Lang, "An overview of metamap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.
- [82] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [83] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019.
- [84] S. Bird, "Nltk: the natural language toolkit," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 2006, pp. 69–72.
- [85] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," 2017, pp. 4444–4451. [Online]. Available: <http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972>

- [86] T. Mikolov, E. Grave, P. Bojanowski, C. Puhersch, and A. Joulin, “Advances in pre-training distributed word representations,” in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
- [87] N. Reimers, “BERT NLI Models,” <https://github.com/UKPLab/sentence-transformers/blob/master/docs/pretrained-models/nli-models.md>, [Online; accessed 20-April-2020].
- [88] C. Shivade, M.-C. de Marneffe, E. Fosler-Lussier, and A. M. Lai, “Extending negex with kernel methods for negation detection in clinical text,” in *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, 2015, pp. 41–46.
- [89] A. W. Chow, M. S. Benninger, I. Brook, J. L. Brozek, E. J. Goldstein, L. A. Hicks, G. A. Pankey, M. Seleznick, G. Volturo, E. R. Wald *et al.*, “Idsa clinical practice guideline for acute bacterial rhinosinusitis in children and adults,” *Clinical Infectious Diseases*, vol. 54, no. 8, pp. e72–e112, 2012.
- [90] B. T. Society, “Scottish intercollegiate guidelines network,” *British Guideline on the management of asthma. Thorax*, vol. 58, no. Suppl 1, pp. i1–94, 2003.
- [91] D. Jurafsky, “Speech and Language Processing,” https://web.stanford.edu/~jurafsky/slp3/slides/4_NB_Jan_10_2021.pdf, [Online; accessed 19-March-2021].
- [92] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, “A survey on text classification: From traditional to deep learning,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 2, pp. 1–41, 2022.
- [93] M. Hofmann and R. Klinkenberg, *RapidMiner: Data mining use cases and business analytics applications*. CRC Press, 2016.
- [94] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. O. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz, “Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals,” *arXiv preprint arXiv:1911.10422*, 2019.

- [95] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.
- [96] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.
- [97] A. Lex, N. Gehlenborg, H. Strobel, R. Vuilleumot, and H. Pfister, "Upset: visualization of intersecting sets," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1983–1992, 2014.
- [98] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [99] M.-H. Hsu, M.-F. Tsai, and H.-H. Chen, "Query expansion with conceptnet and wordnet: An intrinsic comparison," in *Asia Information Retrieval Symposium*. Springer, 2006, pp. 1–13.
- [100] A. D. Association, "2. classification and diagnosis of diabetes: Standards of medical care in diabetes—2020," *Diabetes care*, vol. 43, no. Supplement_1, pp. S14–S31, 2020.
- [101] C. E. UK, "Type 2 diabetes in adults: management," 2015.
- [102] S. I. Guideline Network, "Management of diabetes: a national clinical guideline," *SIGN*, 2010.
- [103] A. Neustein, S. S. Imambi, M. Rodrigues, A. Teixeira, and L. Ferreira, "Application of text mining to biomedical knowledge extraction: analyzing clinical narratives and medical literature," *Text Mining of Web-based Medical Content*, pp. 3–32, 2014.
- [104] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones, R. Forshee, M. Walderhaug, and T. Botsis, "Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review," *Journal of biomedical informatics*, vol. 73, pp. 14–29, 2017.

Acronyms

In alphabetical order:

AI Artificial Intelligence

CA Condition Action

CC Condition Consequances

CDSS Clinical Decision Support System

CPG Clinical Practice Guideline

CTTM Casual Triple Trained Model

EHR Electronic Health Record

KE Knowledge Engineer

ML Machine Learning

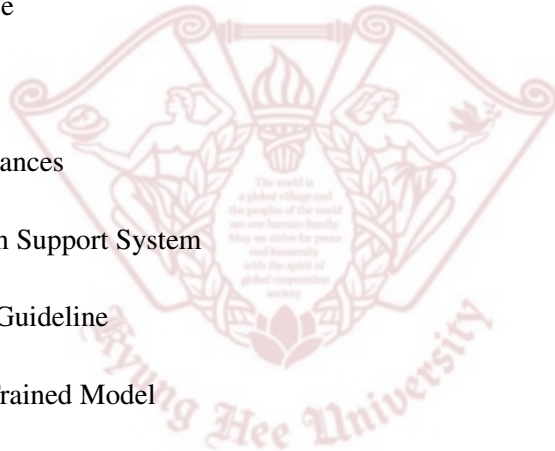
NBC Naive based Classifier

NER Named Entity Recognition

NGT Nominal Group Technique

NLP Natural Lanauge Processing

NC No Condition



NP Noun Phrase

NRS Non Recommendation Sentence

POS Part of Speech

REDEx Regular Expression Discovery Extractor

RS Recommendation Sentence

SVM Support Vector Machine

UMLS Unified Medical Language System



B.1 International Journal Papers [9]

- 1 **Hussain, Musarrat**, Fahad Ahmed Satti, Jamil Hussain, Taqdir Ali, Syed Imran Ali, Hafiz Syed Muhammad Bilal, Gwang Hoon Park, Sungyoung Lee, and TaeChoong Chung. "A practical approach towards causality mining in clinical text using active transfer learning." *Journal of Biomedical Informatics* 123 (2021): 103932.
- 2 **Hussain, Musarrat**, Fahad Ahmed Satti, Syed Imran Ali, Jamil Hussain, Taqdir Ali, Hun-Sung Kim, Kun-Ho Yoon, TaeChoong Chung, and Sungyoung Lee. "Intelligent knowledge consolidation: From data to wisdom." *Knowledge-Based Systems* 234 (2021): 107578.
- 3 **Hussain, Musarrat**, Jamil Hussain, Taqdir Ali, Syed Imran Ali, Hafiz Syed Muhammad Bilal, Sungyoung Lee, and Taechoong Chung. "Text Classification in Clinical Practice Guidelines Using Machine-Learning Assisted Pattern-Based Approach." *Applied Sciences* 11, no. 8 (2021): 3296.
- 4 Satti, Fahad Ahmed, **Musarrat Hussain**, Jamil Hussain, Syed Imran Ali, Taqdir Ali, Hafiz Syed Muhammad Bilal, Taechoong Chung, and Sungyoung Lee. "Unsupervised Semantic Mapping for Healthcare Data Storage Schema." *IEEE Access* 9 (2021): 107267-107278.
- 5 Imran Ali, Syed, Bilal Ali, Jamil Hussain, **Musarrat Hussain**, Fahad Ahmed Satti, Gwang Hoon Park, and Sungyoung Lee. "Cost-sensitive ensemble feature ranking and automatic threshold selection for chronic kidney disease diagnosis." *Applied Sciences* 10, no. 16 (2020): 5663.
- 6 Ali, Syed Imran, Hafiz Syed Muhammad Bilal, **Musarrat Hussain**, Jamil Hussain, Fahad

- Ahmed Satti, Maqbool Hussain, Gwang Hoon Park, Taechoong Chung, and Sungyoung Lee. "Ensemble feature ranking for cost-based non-overlapping groups: A case study of chronic kidney disease diagnosis in developing countries." *IEEE Access* 8 (2020): 215623-215648.
- 7 Bilal, Hafiz Syed Muhammad, Muhammad Bilal Amin, Jamil Hussain, Syed Imran Ali, Muhammad Asif Razzaq, **Musarrat Hussain**, Asim Abbas Turi, Gwang Hoon Park, Sun Moo Kang, and Sungyoung Lee. "Towards user-centric intervention adaptiveness: influencing behavior-context based healthy lifestyle interventions." *IEEE Access* 8 (2020): 177156-177179.
- 8 Ali, Taqdir, Jamil Hussain, Muhammad Bilal Amin, **Musarrat Hussain**, Usman Akhtar, Wajahat Ali Khan, Sungyoung Lee et al. "The Intelligent Medical Platform: A novel dialogue-based platform for health-care services." *Computer* 53, no. 2 (2020): 35-45.
- 9 Kim I, Lee JH, Choi D, Sung-ji Park, Ju-Hee Lee, Sang Min Park, Mina Kim, Hack-Lyoung Kim, Sunki Lee, In Jai Kim, Seonghoon Choi, Jaehun Bang, Bilal Ali, **Musarrat Hussain**, Taqdir Ali, Sungyoung Lee "Rationale design and efficacy of a smartphone application for improving self-awareness of adherence to edoxaban treatment: study protocol for a randomised controlled trial (adhere app)" *BMJ Open* 12.4 (2022): e048777.

B.2 Domestic Journal Papers [4]

- 1 **Musarrat Hussain**, Taqdir Ali, Jamil Hussain, Fahad Ahmed Satti, Usman Akhtar, Jaehun Bang, Taeho Hur, Sun Moo Kang, Byeong Ho Kang, and Sungyoung Lee. "Intelligent Medical Platform: IMP", *The Journal of The Korean Institute of Communication Science*, 37, no. 9(2020): 3-17.
- 2 **Musarrat Hussain**, Maqbool Hussain, Muhammad Afzal and Sungyoung Lee. "Contemporary CDSS Frameworks and A Case Study of Smart CDSS", *The Journal of The Korean Institute of Communication Science*, 35, no. 2(2018): 18-32.
- 3 Jaehun Bang, Taeho Hur, Taqdir Ali, **Musarrat Hussain**, and Sungyoung Lee. "Intelligent-Knowledge Authoring Tool(I-KAT)", *The Journal of The Korean Institute of Communica-*

tion Science, 37, no. 9(2020): 18-37.

- 4 Dong-Ju Choi, Taqdir Ali, Jin Joo Park, **Musarrat Hussain**, and Sungyoung Lee. "Cardiovascular Silos: Intelligent Decision-Making Systems for Heart Failure Diagnosis", The Journal of The Korean Institute of Communication Science, 37, no. 9(2020): 53-60.

B.3 International Conference Papers [10]

- 1 **Hussain, Musarrat**, Jamil Hussain, Taqdir Ali, and Sungyoung Lee. "An Empirical Method of Automatic Pattern Extraction for Clinical Text Classification." In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 5292-5295. IEEE, 2020.
- 2 **Hussain, Musarrat**, Dong-Ju Choi, and Sungyoung Lee. "Semantic based clinical notes mining for factual information extraction." In 2020 International Conference on Information Networking (ICOIN), pp. 46-48. IEEE, 2020.
- 3 **Hussain, Musarrat**, and Sungyoung Lee. "Information extraction from clinical practice guidelines: A step towards guidelines adherence." In International Conference on Ubiquitous Information Management and Communication, pp. 1029-1036. Springer, Cham, 2019.
- 4 **Hussain, Musarrat**, Jamil Hussain, Muhammad Sadiq, Anees Ul Hassan, and Sungyoung Lee. "Recommendation statements identification in clinical practice guidelines using heuristic patterns." In 2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 152-156. IEEE, 2018.
- 5 **Hussain, Musarrat**, Anees Ul Hassan, Muhammad Sadiq, Byeong Ho Kang, and Sungyoung Lee. "Missing information prediction in ripple down rule based clinical decision support system." In International Conference on Smart Homes and Health Telematics, pp. 179-188. Springer, Cham, 2018.
- 6 Satti, Fahad Ahmed, **Musarrat Hussain**, Sungyoung Lee, and TaeChoong Chung. "Significance of Syntactic Type Identification in Embedding Vector based Schema Matching." In

- 2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM), pp. 1-6. IEEE, 2022.
- 7 Satti, Fahad Ahmed, **Musarrat Hussain**, Jamil Hussain, Tae-Seong Kim, Sungyoung Lee, and TaeChoong Chung. "User Stress Modeling through Galvanic Skin Response." In 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM), pp. 1-6. IEEE, 2021.
 - 8 Ansaar, Muhammad Zaki, Jamil Hussain, Asim Abass, **Musarrat Hussain**, and Sungyoung Lee. "User's Emotional eXperience Analysis of Wizard Form Pattern Using Objective and Subjective Measures." In International Conference on Web Engineering, pp. 521-524. Springer, Cham, 2019.
 - 9 Sadiq, Muhammad, Muhammad Bilal Amin, Hafiz Syed Muhammad Bilal, **Musarrat Hussain**, Anees Ul Hassan, and Sungyoung Lee. "LogMap-P: On matching ontologies in parallel." In Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication, pp. 1-5. 2018.
 - 10 Hassan, Anees Ul, Jamil Hussain, **Musarrat Hussain**, Muhammad Sadiq, and Sungyoung Lee. "Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression." In 2017 International Conference on Information and Communication Technology Convergence (ICTC), pp. 138-140. IEEE, 2017.

B.4 Patents [4]

- 1 Sungyoung Lee and **Musarrat Hussain**, "Methods and systems for converting clinical practice guidelines into computer-interpretable models" *Japan Patent Office*, Registered: 6974878 (2021.11.09).
- 2 Sungyoung Lee and **Musarrat Hussain**, "System and method for converting clinical practice guideline to computer interpretable model." *Korean Intellectual Property Office*, Registered: 1020190082070 (2020.05.31).

- 3 Sungyoung Lee and **Musarrat Hussain**, "System and method for converting clinical practice guideline to computer interpretable model." U.S. Patent Application No. 20210012896A1 Applied on: 2020.05.18.
- 4 Sungyoung Lee and **Musarrat Hussain** "An intelligent system for patient health summary generation" *Korean Intellectual Property Office*, Application No. 10-2020-0148833, Applied on: 2020.11.09.

