KYUNG HEE UNIVERSITY

Department of Computer Science & Engineering, KHU, South Korea

PhD Dissertation Presentation

Knowledge Extraction from Unstructured Clinical Text using Active Transfer Learning Approach

Musarrat Hussain

musarrat.hussain@oslab.khu.ac.kr

Advisors: **Prof. Sungyoung Lee, PhD**
Prof. TaeChoong Chung, PhD

Ubiquitous Computing Laboratory
Kyung Hee University, Korea

11 May 2022

PRESENTATION AGENDA

INTRODUCTION
- Background
- Motivation
- Problem Statement
- Taxonomy
- Related Work

PROPOSED SOLUTION
- Solution 1: Clinical Text Classification
- Solution 2: Causality Mining
- Solution 3: Rules Generation

EXPERIMENTS & RESULTS
- Dataset
- Experimental Setup
- Results & Discussion

CONCLUSION

Background

Unstructured Clinical Text

- Artificial Intelligence (AI) has the potential to solve many important problems impacting patients, providers and health systems.[1]
- Nearly 80% of clinical information in electronic health records (EHRs) is “Unstructured” and in a format that health information technology systems cannot use. [2, 3]
- The unstructured clinical text stored in the EHR systems is among the most significant barriers to healthcare quality improvement. [4]

Motivation

[23 - 25]

Patient Profile Data

- Personal data
- Demographics
- Social aspects

[23 - 25]

General Medical Knowledge

- Disease information
- Symptoms
- Available medications

02

01

03

Clinician

[23 - 25]

Medical Procedures

- Define clinical actions
- Procedures
- Clinical context

CPG

Clinical Notes

Clinical Protocols

Online Research

Knowledge required for making clinical decisions

NLP based Text Processing

- A plethora of **valuable** clinical data is available in **unstructured textual** format.
- **Manual processing** of the available textual data is **challenging** and **resource** intensive.[5]
- This **necessitates** an **automatic** or semi-automatic natural language processing solution. [6]

CPG: Clinical Practice Guidelines
NLP: Natural Language Processing

Problem Statement

Problem statement

Clinical text withholds **implicit knowledge**[8], which provides a rich **source** for applying and enhancing **clinical practices**[9, 10]. Identification of a machine readable representation of this knowledge necessitates a **stable**, **scalable**, and **semi-automatic** mechanism[30,31].

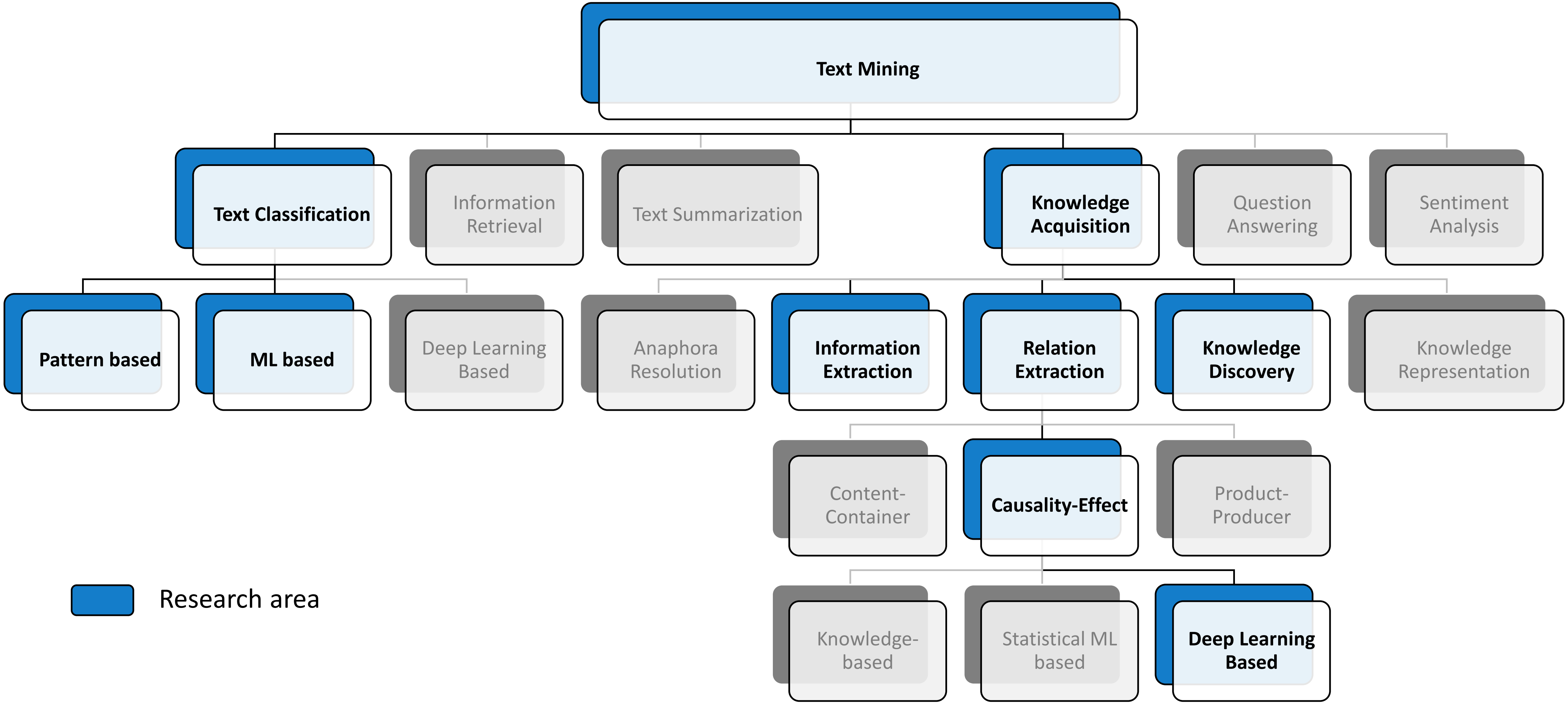
Goal

Aims to identify and extract **clinical knowledge** from unstructured clinical resources and transform it into **machine-interpretable** knowledge.

Challenges

- Challenge 1: Classify recommendation and non-recommendation statements in the text [11, 12, 49]
- Challenge 2: Identify and extract clinical concepts and their relationships[13, 14]
- Challenge 3: Generate Knowledge from the extracted concepts and their relationships

Research Taxonomy



[26, 27]

Related Work

Literature Survey for Clinical Text Classification, Concepts & Relation Extraction

	Research	Method & Advantages	Limitations
Sentence Classification	[Bui2014][15]	o Proposed an **automatic** regex extraction algorithm (RED) o RED produces **two regular expressions**, with and without distance control.	The extracted patterns utilize **exact terms** that lack generalization and suffer from tokens' **polysemy** and **synonyms**.
	[Morid2016][42]	o Applied a Kernel-based Bayes Network classifier o Evaluated **various** features including UMLS concepts, **semantic** predications, **patient** population and **cue** words.	o This approach was **tuned for treatment** recommendations and need **substantial tuning** for other type of recommendations such as diagnosis. o Experiments were conducted on **small dataset**
	[VanDam2017][43]	o Evaluated various combinations of feature set o Used KNN, Naïve Bayes, and SVM for content classification	o The models resulted **lower precision** on other unseen dataset. o The models **face difficulties** in classifying **short**, **diet** related, and **general** healthcare related sentences.
	[Hematialam2021][17]	POS tags make the approach more **generalized** and can be applied to various domains.	The POS based approach **lacks** Semantic information.
Concepts & Relation Extraction (Causality Mining)	[Doan2019][19]	o Used a **huge dataset** of 24 Million tweets o Targeted Stress, insomnia, and headache **domains**.	Created a set of **six patterns** to identify cause-effect relationship, which may not be enough to **handle all casual** relations.
	[Pawar2021][44]	o Used **unsupervised** approach to discover causal **triggers**. o Set l**inguistic rules** for cause-effect arguments of the triggers.	o **Manually** designed the classification **rules**. o The designed rules lacks **generalizations**
	[De2017][45]	o Trained SVM with Knowledge based features (KF) and CNN with KF, Tri-section, position embedding and their combination o Causality direction was also handled	The CNN **performed better** than SVM on the same dataset. However, the trained CNN model showed **tendencies** towards **overfitting** as its performance **decreased** on other corpus.
	[AnNing2019][20]	o Initiate **seed triggers** that lifts training data requirements for ML. o Introduced **word vector** approach for causality mining	Embedding **only causal trigger** (verb) **lacks semantic** and leads to incorrect causality detection

Related Work

Literature Survey for Concept's Value Extraction

Rule Generation (Concepts Value Extraction)

Research	Method & Advantages	Limitation
[Redd2016][46]	o Extend **REDEx** algorithm by improving pattern **generalization** for **METs Value** Extraction o **Replaced** pattern concepts with its **equivalent length pattern** for generalized regular expression.	o Required **detailed training data** including Before Label Segment (BLS), Label Segment (LS), and After Label Segment (ALS) o Considered **token length** for generalization, while token can have **diverse length alternatives** that will result in errors.
[Zheng2018][22]	o Formulated value extraction as **sequence tagging** task similar to named entity recognition. o Used **LSTM** for **semantic and context** and **CRF** to enforce tagging **consistency** and extract **cohesive chunks** of attributes values.	o Uses self-attention to capture the important tokens in the title, but treats attribute only as a type. o Neglects **attribute semantic** information. [28]
[Cai2019][47]	o Developed NLP tool (EXTEND) for extracting **vital signs** and cardiac ejection fractions (**EF**) values. o Utilized **dictionary** for concepts detection and **rules** for values extraction and validation.	o Dictionary and rules based solutions **lacks generalization**, and required **human efforts** to extend underlying dictionary and rules for accommodating new concepts.

Challenge 1: Limitations of existing work

- Mostly used **handcrafted patterns** which **lack generalization** and require intensive human efforts and time.
- ML models produce lower accuracies, while DL models lack **large training data** requirements.

Challenge 2: Limitations of existing work

- Rule based approaches can not handle all cause-effect relations due to **data Sparsity and diversity.**
- Word embedding applied on **only triggers** (verb) lacks **contextual information**.

Challenge 3: Limitations of existing work

- Most of the researchers used **dictionary and rule base** solutions, which lack generalization, and extensive human efforts are required to extend solutions to new concepts or domains.

Challenges and Proposed Solutions

Challenges	Proposed Solutions	Objectives
Challenge: 1 Classify recommendation and non-recommendation statements in clinical text.	**Solution 1** **Sentence Classification** Automatic and Machine Learning assisted pattern based approach for sentence classification.	Find clinical statements which contain recommendations intended to optimize patient care.
Challenge: 2 Identify and extract clinical concepts and their relationships.	**Solution 2** **Causality Mining** POS based clinical concepts identification and BERT based causality mining for relations extraction.	Find relevant concepts and their relationships
Challenge: 3 Generate knowledge from the extracted concepts and their relations.	**Solution 3** **Rule Generation** Cause-effect mapping to condition-actions and their quantitative terms identification.	Transform the identified concepts and their relationships into machine interpretable format such as production rule.

Goal

To transform unstructured clinical text into machine understandable and transparent knowledge

Proposed Methodology

Abstract View

Research Papers (7)

- Published **3** journal papers
- Published **4** conference papers

Clinical Text

Solution 1: Sentence Classification

- Load Clinical Text
- Text Preprocessing
- Sentence Classification

Recom. Sentences

Solution 2: Causality Mining

- Triples Generation
- Triples Embedding
- Triples Evaluation
- Causal Relation Extraction

Extended Causal Triples
<Noun Phrase, Verb Phrase, Noun Phrase, Metadata>

Solution 3: Rules Generation

- Condition(s) Identification
- Action Identification
- Rule Generation

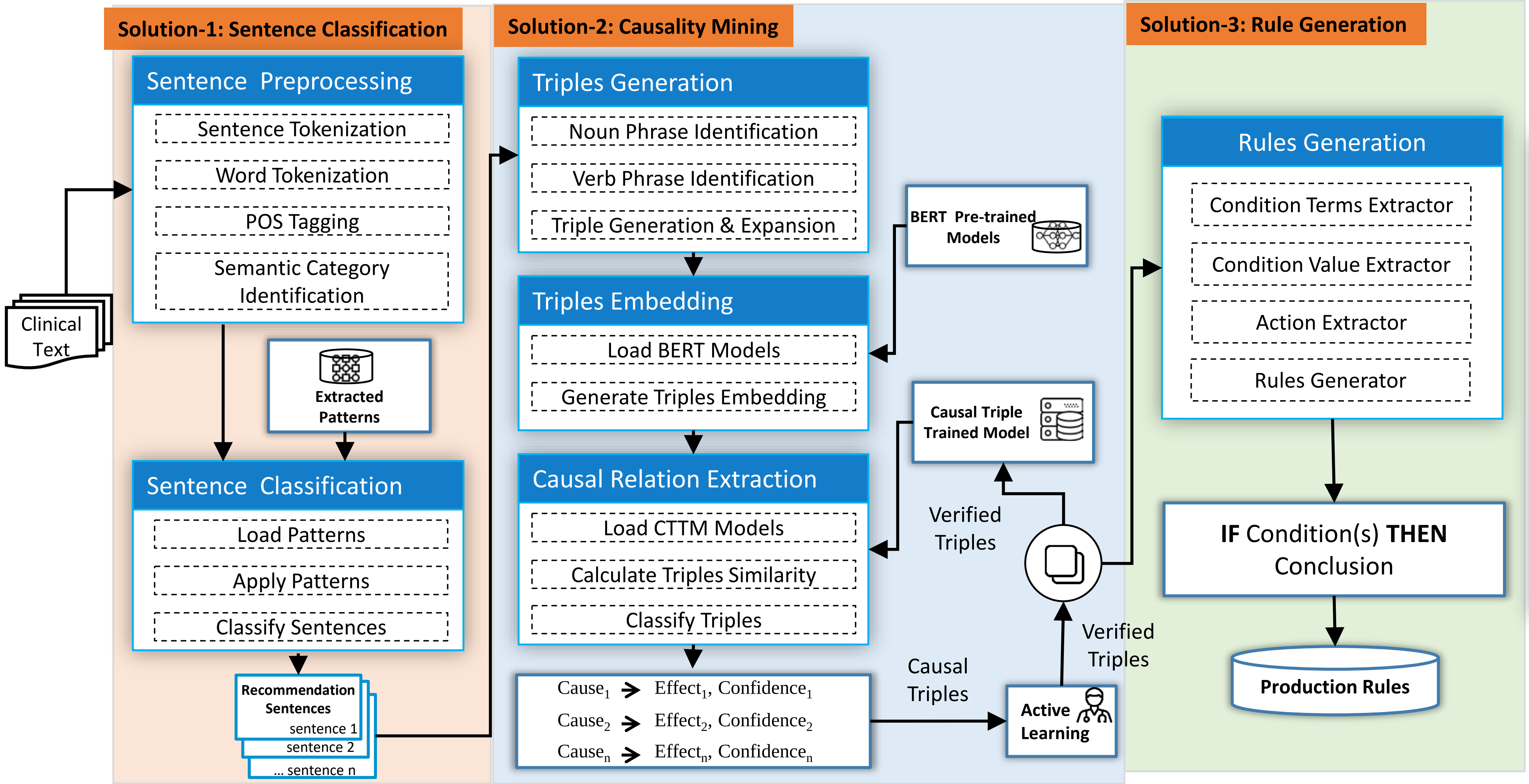
Production Rules

Patents Registration (4)

- Registered **1** international and **1** local patent
- Applied **1** international and **1** local patent

Proposed Methodology

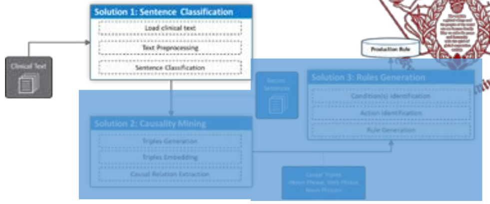
Detail Workflow



Contributions

- Automatic and **Machine Assisted** patterns for sentence classification
- Active transfer learning and **BERT Multi-Model** based causality mining approach
- Transformation of Cause-Effect relation to production rules.

Solution 1: Sentence Classification



Possible Approaches

Pattern based Approaches

- Utilize sequence and structure of terms for content classification

Advantages

- ❑ **Transparent** & easy to use[36]
- ❑ Still **mostly used** approach in real applications[32,35]
- ❑ Can work on **imbalance data** [35]

Disadvantages

- ➢ Required human efforts and time for pattern extraction
- ➢ Suffer from polysemy and synonyms[33]

Traditional ML Approaches

- Take labeled data as input
- Features extraction and selection
- Model training

Advantages

- ❑ Can work on **small data**[34]
- ❑ Easier to **interpret**.
- ❑ Financially and computationally **affordable**

Disadvantages

- ➢ Required feature engineering
- ➢ Need to re-trained on new labeled data.

Deep Learning Approaches

- Take labeled data as input
- Automatic feature engineering
- Model training

Advantages

- ❑ **No feature** engineering required
- ❑ Suitable for **large data**
- ❑ **Adaptable** and transferable

Disadvantages

- ➢ Required huge labeled data
- ➢ Non interpretable
- ➢ Cannot be tuned manually[36]

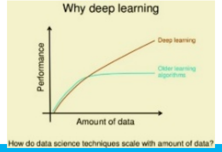


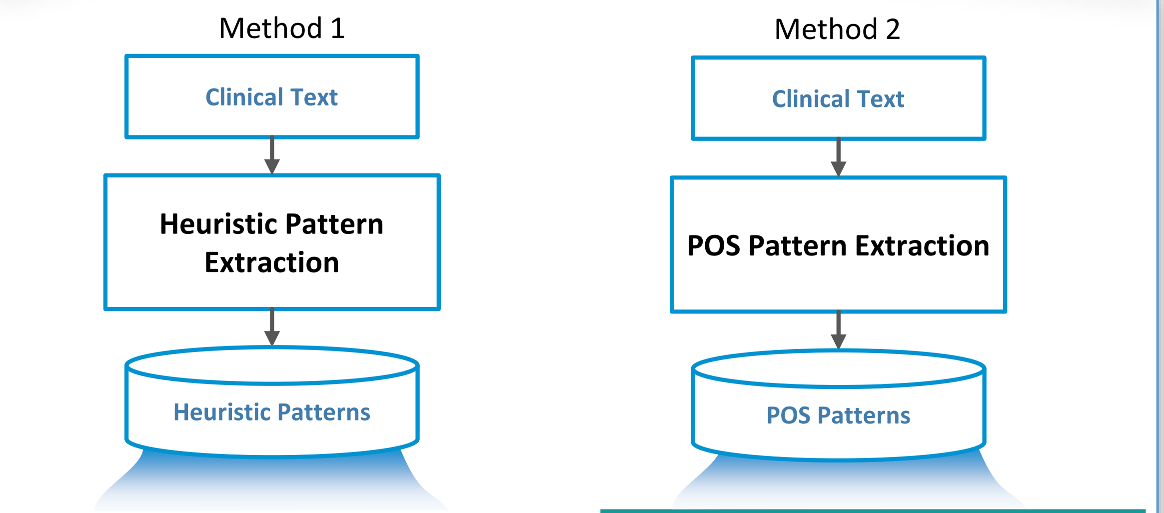
Image Source: https://towardsdatascience.com/deep-learning-vs-classical-machine-learning-9a42c6d48aa

Solution 1: Sentence Classification

Pattern based Approach As-Is vs To-Be

Existing Methods

[15, 16, 17, 21, 36]



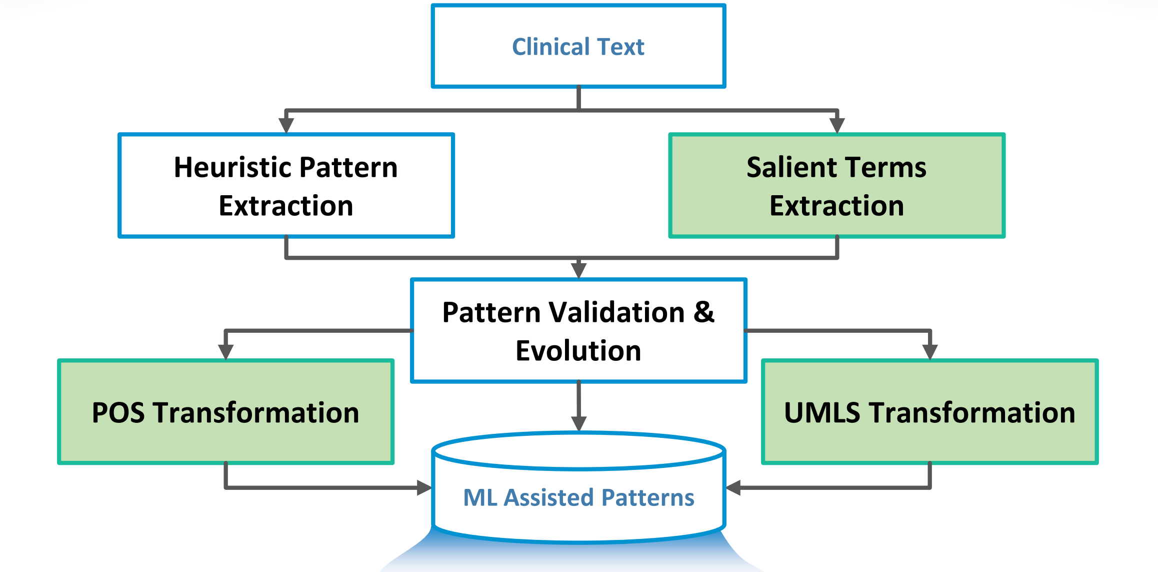
Benefits
- Heuristic pattern extraction is mostly used approach
- Produces high accuracy

Limitations
- Lacks generalization

Benefits
- Generalizable patterns

Limitations
- Requires human efforts and time
- Lacks semantic information

Proposed Method

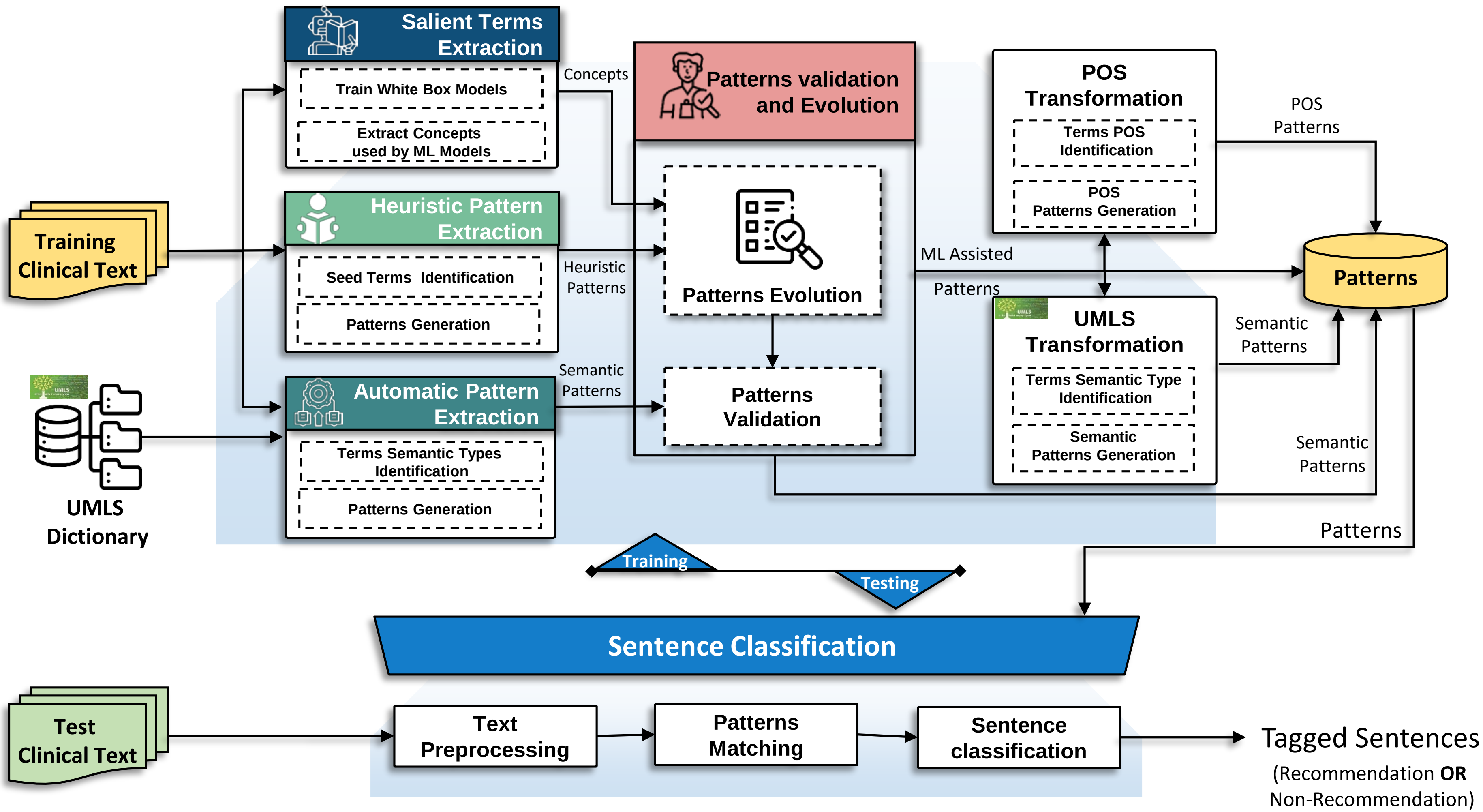


Benefits
- Handles domain dependent (heuristic) and independent (POS, Semantic type) features.
- Uses ML assistance for pattern extraction

- The terms **utilized by ML models** for sentence **classification** is considered as **Salient Terms**.

Solution 1: Sentence Classification

Machine Learning Assisted Pattern Extraction Process Flow



Why Salient terms

- Salient terms are utilized to reflect machine identified insights in the patterns.

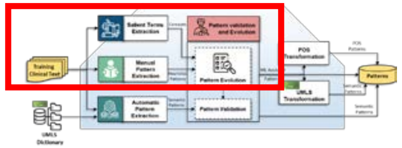
Contributions

- ✓ **ML Assisted pattern** extraction methodology, pattern generalization with **POS** and **UMLS** Semantic Concepts.
- ✓ **An automatic** pattern extraction **algorithm**

Benefits

- ✓ The extracted patterns reflect **human expertise** as well as **ML based** insights**.**
- ✓ The extracted patterns are more **generalized** as they also consider **POS** and **UMLS** concepts.

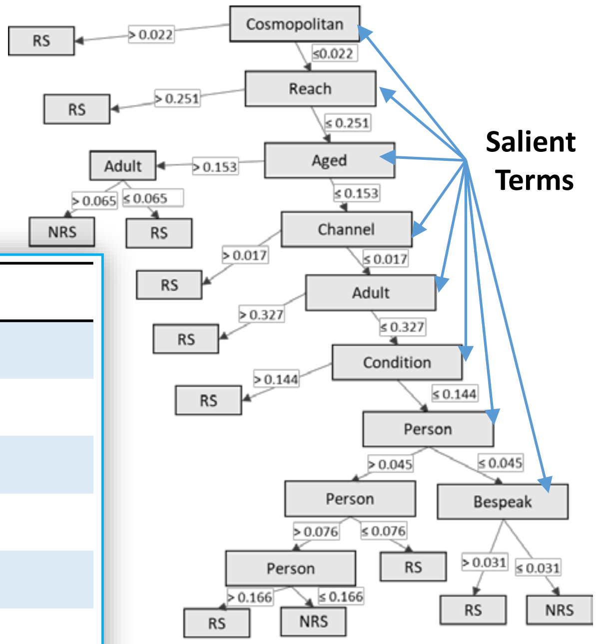
3
Hussain, Musarrat, et al. "Text Classification in Clinical Practice Guidelines Using Machine-Learning Assisted Pattern-Based Approach." *Applied Sciences* 11.8 (2021): 3296.
Hussain, Musarrat, et al. "Recommendation statements identification in clinical practice guidelines using heuristic patterns." *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*. IEEE, 2018.
Hussain, Musarrat, et al. "An Empirical Method of Automatic Pattern Extraction for Clinical Text Classification." *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020.



Solution 1: Sentence Classification

Machine Learning Assisted Pattern Extraction

DT Model



Heuristic Pattern Extraction

Clinical Text

Nominal Group Technique (NGT) [41]

Extract Patterns Individually

Combine Extracted Patterns

Rank Each Pattern from 0 – 5

$$Pattern_{Score} = \sum_{i=1}^{5} Pattern_{Score}(KEi)$$

Select if $Pattern_{Score}$ >= 15

Example Pattern: *.*(patient(s)?)?with (disease).**

Salient Terms Extraction

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

Modify extracted patterns by utilizing the ML considered terms (Salient Terms)

ML Assisted Patterns

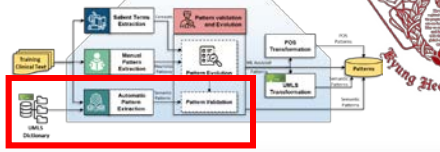
POS Patterns Generation

UMLS Patterns Generation

Example Pattern: *.*(patient(s)?|adult|(population group))?with (disease).**

1 Hussain, Musarrat, et al. "Text Classification in Clinical Practice Guidelines Using Machine-Learning Assisted Pattern-Based Approach." *Applied Sciences* 11.8 (2021): 3296.

Solution 1: Sentence Classification



Automatic Pattern Extraction

Automatic Pattern Extraction Algorithm

inputs: Training Corpus D, $UMLS$, Concept Threshold CT, Context Window Threshold CWT
output: Patterns $P = [p_1, p_2, p_3, \ldots p_n]$

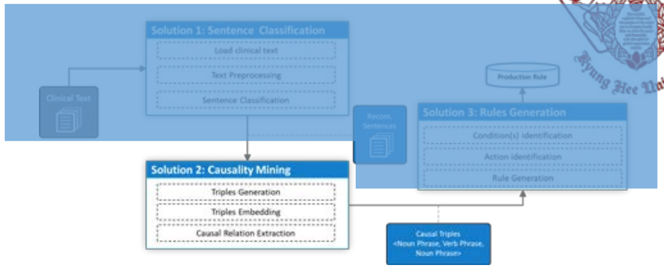
1. **P ← []**
2. **for** each document $d \in D$ **do**
3. Concepts $C \leftarrow []$
4. Sentences $S \leftarrow$ Sent_tokenize(d)
5. **for** each sentence $s_i \in S$ **do**
6. $s_i \leftarrow s_i$.lower()
7. Sentence concepts $SC \leftarrow$ word_tokenize(s_i)
8. $SC \leftarrow [w_j$ **for** word w_j **in** SC **if** not w_j **in** stopwords.words()]
9. C.append(SC)
10. Concept Semantics $CS \leftarrow []$
11. **for** concept c_i in C **do**
12. Concept Semantics $CS \leftarrow$ token_Semantics(c_i, $UMLS$)
13. Unique Concepts $UC \leftarrow$ Counter(C).keys()
14. Candidate Concepts $CC \leftarrow []$
15. **for** concept c_i in UC **do**
16. **if** count(c_i) in C > CT **do**
17. CC.append(c_i)
18. Context Window $CW \leftarrow []$
19. **for** concept c_i in CC **do**
20. context Window $CW.add[c_{i-n}, \ldots, c_{i-2}, c_{i-1}, c_i, c_{i+1}, c_{i+2}, \ldots, c_{i+n}])$ where $i = 1, 2, 3, \ldots, n$, $c_{i-1}, c_{i-2}, \ldots c_{i-n}$ and $c_{i+1}, c_{i+2}, \ldots, c_{i+n}$ represents the preceding and succeeding concepts, respectively.
21. **for** cw in CW **do**
22. **if** $count(cw) > CWT$ **do**
23. $P.add$(generate_pattern(cw)#.*(Quantitiative Concept).*(functional Concept).*(Disease or Syndrome)
24. **return** P

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']
Filtered 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).*(Idea or Concept).*]

1 Hussain, Musarrat, et al. "An Empirical Method of Automatic Pattern Extraction for Clinical Text Classification." *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020.

Solution 2: Causality Mining

Abstract View



Solution 1 output

Recommendation Sentences

Preprocessing

- ❑ Clinical documents contain many defacto linguistic elements.
- ❑ Preprocessing cleans and prepares the text for causal phrase detection and extraction

Model Development

- ❑ Expands causal phrases to include related Nouns and Verbs.
- ❑ Converts causal phrases into embedding vectors using BERT NLI models.
- ❑ Amalgamate BERT NLI models into one Causal Trigger Trained Model.

Causality Mining

- ❑ Converts candidate causal phrases into embedding vectors using BERT NLI models.
- ❑ Matches each candidate embedding vector with CTTM.
- ❑ Select all candidate vectors as casual which have been classified at least once.

Feedback

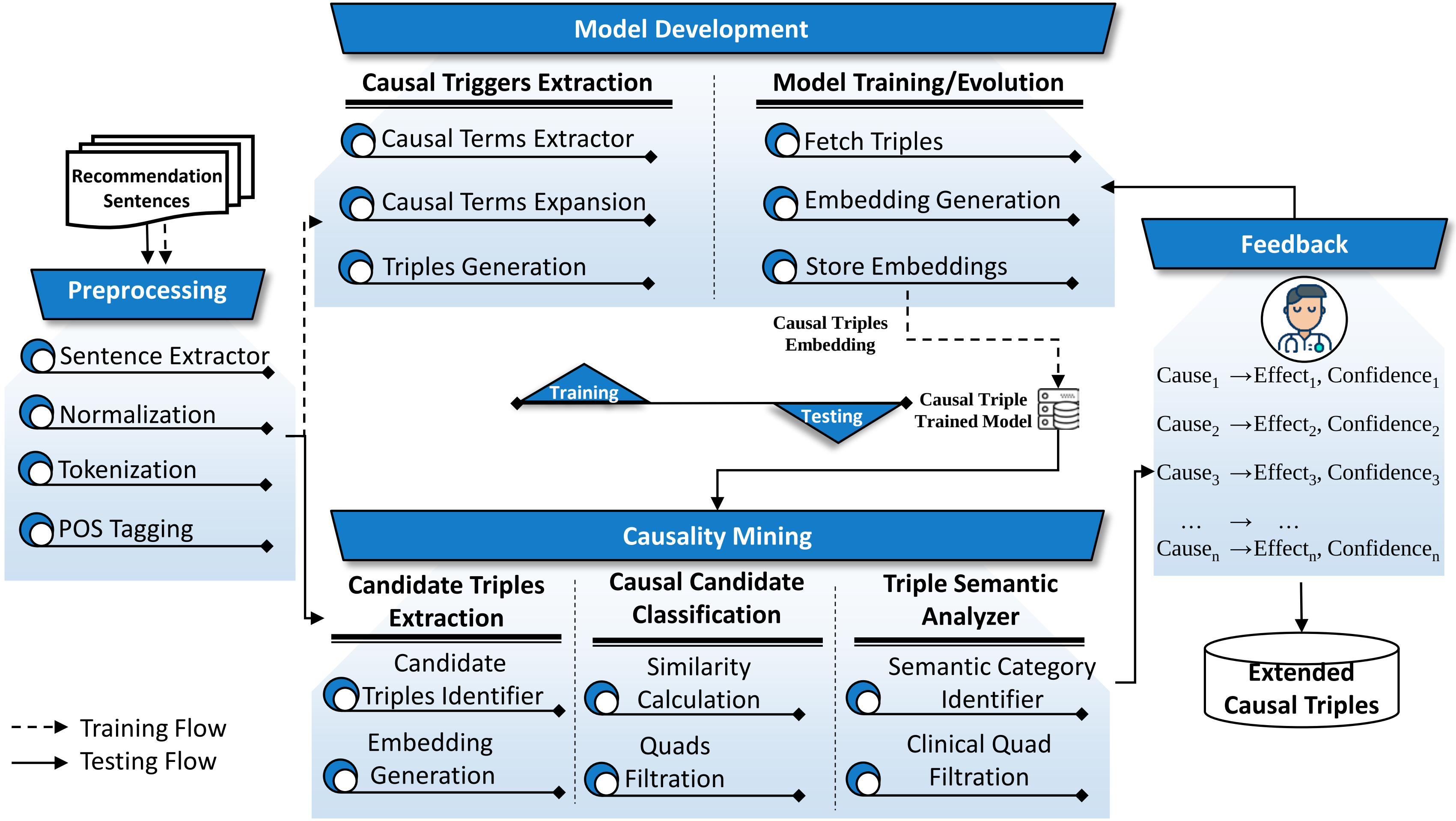
- ❑ Incorporate domain expert feedback for causal tagged triples to evolve CTTM.

Extended Causal Triples

1 Hussain, Musarrat, et al. "A practical approach towards causality mining in clinical text using active transfer learning" Journal of Biomedical Informatics (2021): 103932.

Solution 2: Causality Mining

Detailed Process Flow



Contributions

- **BERT** based **multi-model** approach for distinguishing causal and non-causal triples.
- Incorporating **expert feedback (active learning)** to maximize the effectiveness of causality mining**.**

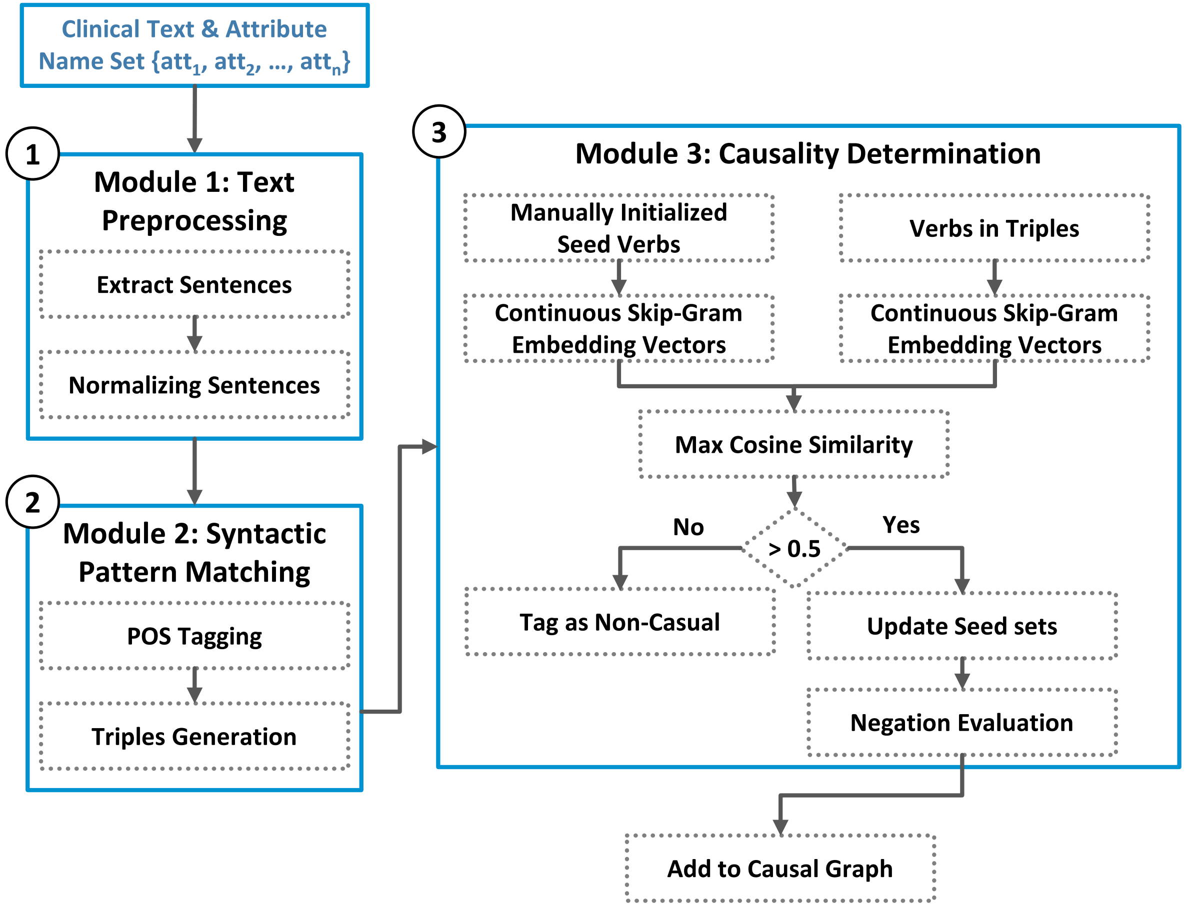
Benefits

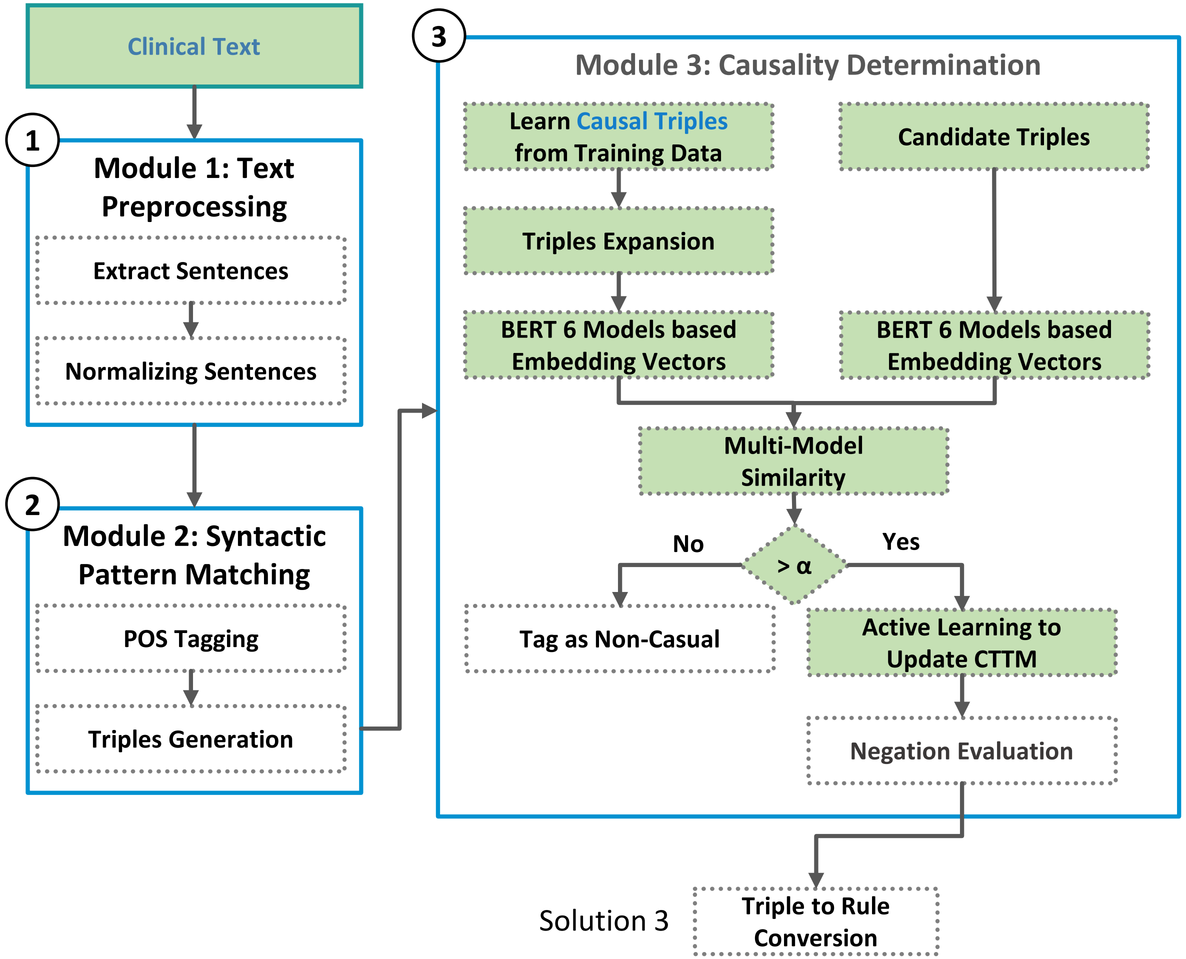
- The BERT based multi-model approach **covers a broader** spectrum compared to an individual model**.**
- **Active learning** enables our model (CTTM) to **update itself** hence providing better results upon subsequent runs.

Hussain, Musarrat, et al. "A practical approach towards causality mining in clinical text using active transfer learning" Journal of Biomedical Informatics (2021): 103932.

Solution 2: Causality Mining

Existing[20] vs Proposed Approach (AS-IS vs TO-BE)

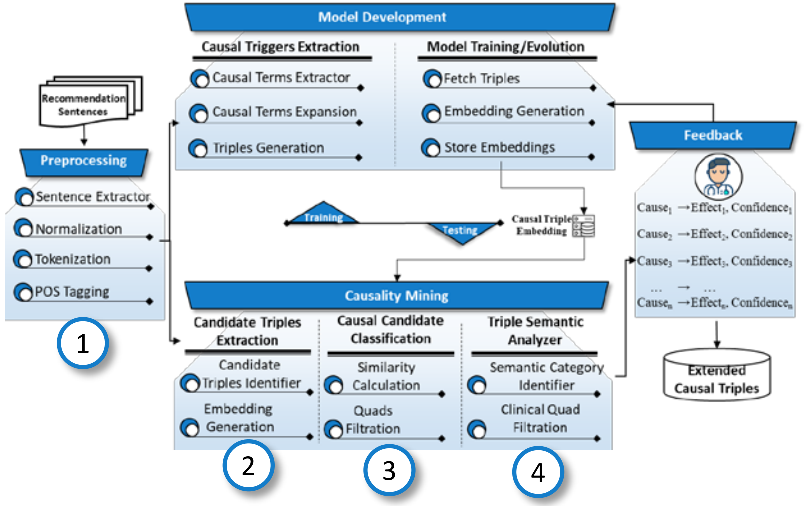




[20] An, Ning, et al. "Extracting causal relations from the literature with word vector mapping." Computers in biology and medicine 115 (2019): 103524.

Solution 2: Causality Mining

Proposed Causality Mining Algorithm



Algorithm 2 : Bert based Multi-Model Causality Mining

inputs: Clinical Documents *D, CTTM*
output: Causal Medical Quad *MQ*

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}
2. **for** each document $d \in D$ **do**
3. Triples $T \leftarrow$ []
4. Sentences $S \leftarrow$ sent_tokenize(d)
5. **for** each sentence $s \in S$ **do**
6. $s \leftarrow$ remove_words_in_brackets(s)
7. $s \leftarrow$ replace_abbreviations(s) 1
8. $s \leftarrow$ nomalize(s)
9. $tokens \leftarrow$ word_tokenize(s)
10. $pos\_tokens \leftarrow$ pos_tag($tokens$)
11. T.append(generate_triples(pos_tokens) # <NP, VP, NP>
12. Causal Quads $CQ \leftarrow$ []
13. **for** triple t in T **do** 2
14. **for** model m in M **do**
15. embedding vector $ev \leftarrow$ embed(t, m)
16. $similarity \leftarrow$ max(similarity(ev, m, $CTTM$)) # Cosine Similarity
17. **if** $similarity > m_\alpha$ **do** 3
18. CQ.append(<$NP, VP, NP, Similarity$>)
19. Medical Quads MQ []
20. **for** cq in CQ **do**
21. $concep_1 \leftarrow$ get_concept(cq, 1)
22. $category_1 \leftarrow$ get_concept_category($concept_1$) # UMLS
23. $concep_2 \leftarrow$ get_concept(cq, 2) 4
24. $category_2 \leftarrow$ get_concept_category($concept_2$) # UMLS
25. **if** $category_1$!= Null AND $category_2$!= Null **do**
26. MQ.append(cq)
27. **return** MQ

Solution 2: Causality Mining

Causal Triples & Candidate Triples Extraction → Example

Step 1. Example Sentence

Celiac <e1>disease</e1> is an inflammatory disorder of the upper small intestine triggered by the <e2>ingestion</e2> of wheat, rye, barley, and possibly oat products.

Step 2. Part-of-Speech Tags

JJ NN NN NN VBZ DT JJ NN IN DT JJ JJ NN VBN IN DT NN NN NN IN NN , NN , NN , CC RB NN NNS .

Celiac <e1>disease</e1> is an inflammatory disorder of the upper small intestine triggered by the <e2> ingestion</e2> of wheat, rye, barley, and possibly oat products.

Step 3. Causal Entities

JJ NN NN NN VBZ DT JJ NN IN DT JJ JJ NN VBN IN DT NN NN NN IN NN , NN , NN , CC RB NN NNS .

Celiac <e1 disease /e1> is an inflammatory disorder of the upper small intestine riggered by the <e2> ingestion /e2> of wheat, rye, barley, and possibly oat products.

Step 4. Causal Elements

disease, is, triggered, ingestion

Step 5. Expansions

disease: diseases, infection, Disease, cancer, infections, incurable_disease, mosquito_borne_disease, … **is:** was, isn'ta, seems, Is, becomes, appears, remains, is, makes, isn_`_t **triggered:** sparked, triggering, precipitated, spurred, prompted, provoked, …, **Ingestion:** ingesting, ingested, ingest, excretion, toxicity, inhalation, …

Step 6. Causal Tipples

<disease, triggered, ingestion>, <infection, triggered, ingestion>, < disease, is, inflammatory disorder >, … , <cancer, is, inflammatory disorder>, <infections, triggered, ingestion>, …

Step 7. Causal Phrases

['disease triggered ingestion', 'infection triggered ingestion', 'disease is inflammatory disorder', … , 'cancer is inflammatory disorder', 'infections triggered ingestion', …]

Step 8. Causal Embedding

[[0.6835734 -0.15927038 0.66193146 … -0.64258146 -0.6853177 0.23629631] [0.65068644 0.07523703 0.64336455 … -0.61406326 -0.48797414 0.32147104] … [0.15254696 -0.14549486 0.65573776 … -0.6533637 -0.4773264 0.22631234] [0.15162535 0.2504582 0.1377482 … -0.8134055 0.07412582 0.6503178] [0.7318822 -0.06919297 0.63599765 … -0.48420542 -0.44523978 0.3177406]…]

Step 1. Example Sentence

Germs are microscopic organisms that cause sickness or disease.

Step 2. Part-of-Speech Tags

NNS VBP JJ NNS WDT VBP NN CC NN .

Germs are microscopic organisms that cause sickness or disease.

Step 3. Candidate Triples

<germs, are, microscopic organisms>, <germs, are, sickness>, <germs, are, disease>, < microscopic organisms, cause, sickness> , < microscopic organisms, cause, disease>

Step 4. Candidate Phrases

['germs are microscopic organisms', 'germs are sickness', 'germs are disease', 'microscopic organisms cause sickness', 'microscopic organisms cause disease']

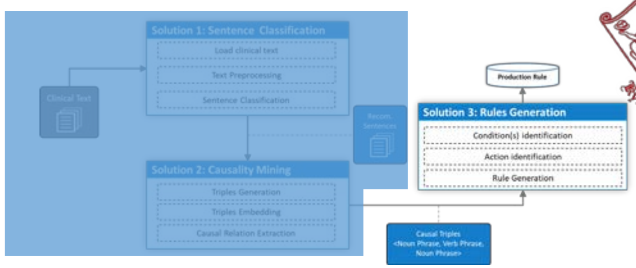
Step 5. Candidate Embedding

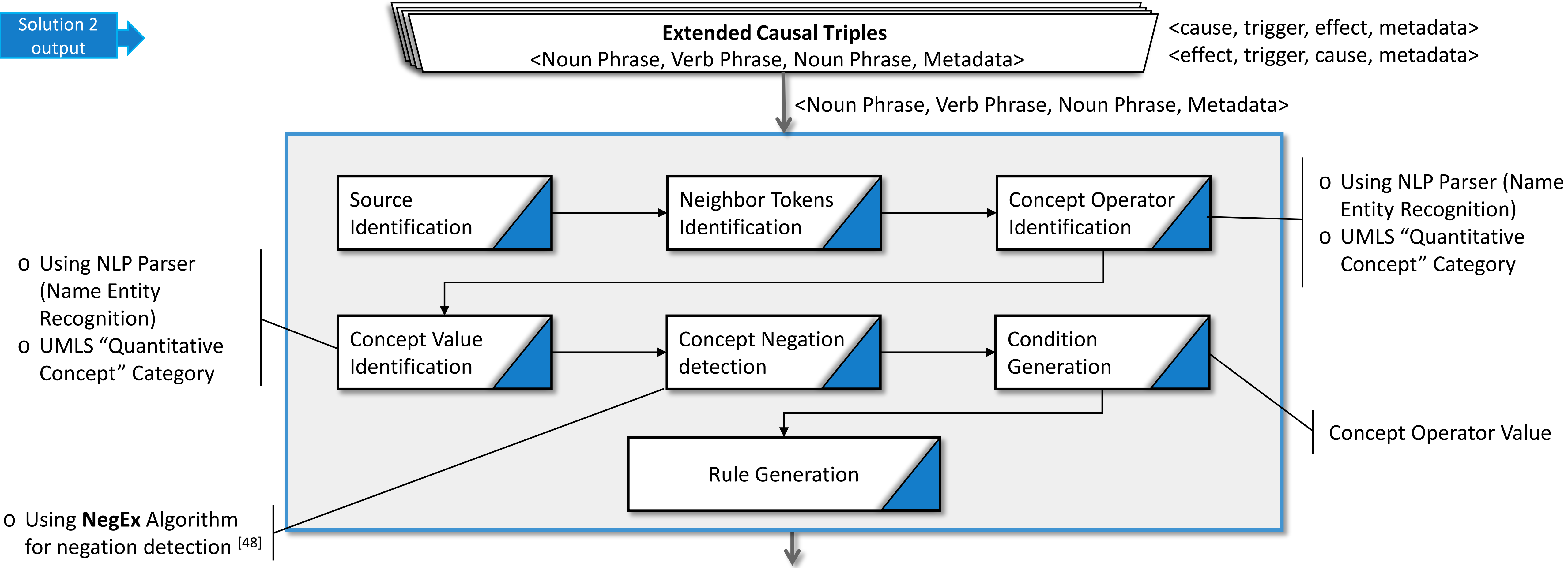
[[0.28555384 0.13149273 0.4754272 … 0.4595495 0.09657189 0.7386813] [0.30529425 0.16654165 1.122178 … -0.7587529 -0.10104396 -0.12151804] [0.45780852 -0.1266355 0.81564075 … -0.4187329 -0.6402518 0.15717138] [0.64769596 0.41299227 0.60672396 … -0.27091768 0.0196734 0.35152712] [0.6362296 0.03878834 0.40926114 … -0.3737266 -0.38633415 0.50904876]]

$$Similarity(A,B) = \frac{A.B}{||A||\ X\ ||B||}$$

$$Candidate_j = \begin{cases} Causal, & \max(similarity(candidate_j, causalTriples\ T)) \geq \alpha \\ Non-casual, & Otherwise \end{cases}$$

Solution 3: Rules Generation





IF *Condition-Key Operator Condition-Value* **THEN** *Conclusion*

Hussain, Musarrat, Dong-Ju Choi, and Sungyoung Lee. "Semantic based clinical notes mining for factual information extraction." 2020 International Conference on Information Networking (ICOIN). IEEE, 2020.

Solution 3: Rules Generation

Triple to Rule Conversion → **Example**

Extended Causal Triples
<Noun Phrase, Verb Phrase, Noun Phrase, Metadata>

<cause, trigger, effect, metadata>
<effect, trigger, cause, metadata>

<cause, trigger, effect>
<plasma glucose, diagnosed, diabetes, metadata>

01 **Source Sentence Identification:**
Patient with hyperglycemic crisis and a **plasma glucose** at or above 200, should be diagnosed with diabetes

02 **Neighbor Tokens Identification:**
Patient with hyperglycemic crisis and a **plasma glucose** at or above 200, should be diagnosed with diabetes

03 **Concept Operator Identification:**
Patient with hyperglycemic crisis and a **plasma glucose** **at or above** 200, should be diagnosed with diabetes

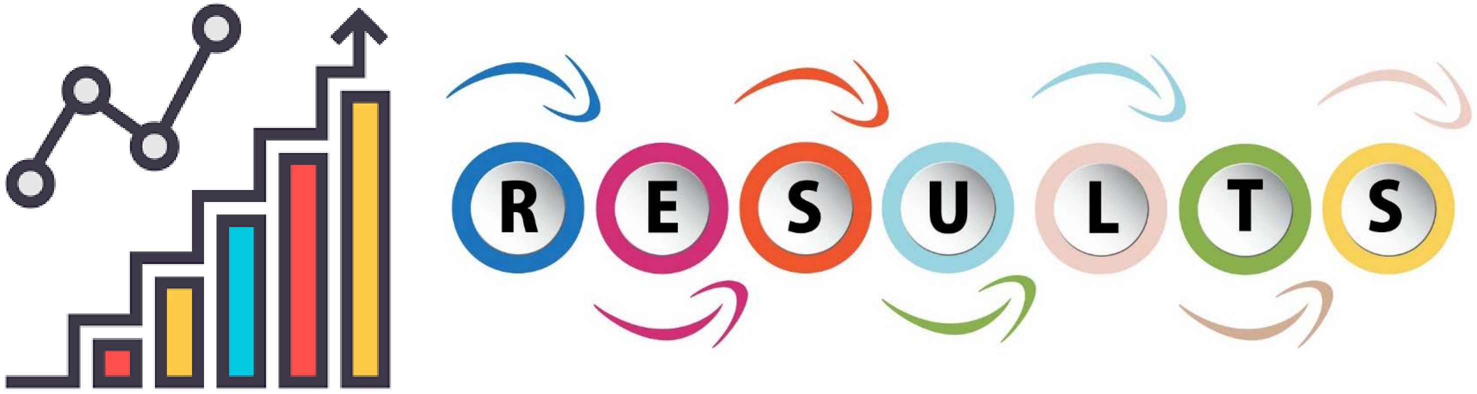
04 **Concept Value Identification:**
Patient with hyperglycemic crisis and a **plasma glucose** **at or above 200**, should be diagnosed with diabetes

05 **Concept Negation Detection:**
Patient with hyperglycemic crisis and a **plasma glucose** **at or above 200**, should be diagnosed with diabetes

06 **Rule Generation:**
IF plasma glucose >= 200 THEN diabetes

IF Cause Operator Value THEN effect

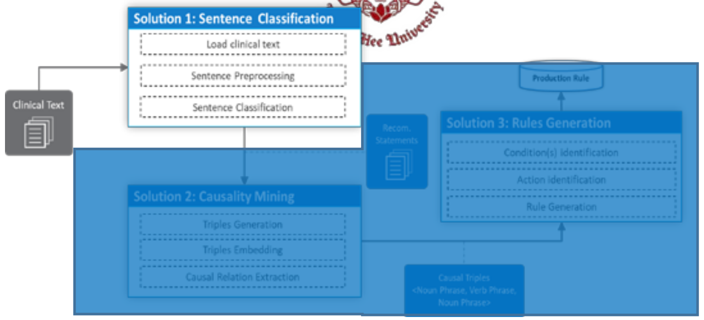
IF plasma glucose >= 200 **THEN** diabetes



RESULTS AND EVALUATION

Experimental Results

Solution 1: Machine Learning Assisted Pattern based Approach



Experimental Setup

Guideline	Total Sentences	Recommendation Sentences	Non-Recommendation Sentences
Hypertension	278	78(28.06%)	200(71.94%)

Split

Training Set

TS	RS	NRS
195	58	137

Test Set

TS	RS	NRS
83	20	63

Heuristic Patterns

<table>
<tr><th>No</th><th>Patterns</th></tr>
<tr><td>1</td><td>.*(add|reomve)(.*)drug.*</td></tr>
<tr><td>2</td><td>.*(recommend(ed)?) treatment.*</td></tr>
<tr><td>3</td><td>.*to improve.*</td></tr>
<tr><td>4</td><td>.*(increase|decrease) .*dose.*</td></tr>
<tr><td>5</td><td>.*treatment (should|with|to).*</td></tr>
<tr><td>6</td><td>.*Recommendation \d+\s+:.*</td></tr>
<tr><td>7</td><td>.*should (include|continue).*</td></tr>
<tr><td>8</td><td>.*(dis)?continue(e|ed|ing|ation).*</td></tr>
<tr><td>9</td><td>.*regardless of.*</td></tr>
<tr><td>10</td><td>.*(patient(s)?)?with (disease).*</td></tr>
</table>

ML Extracted Salient Terms

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

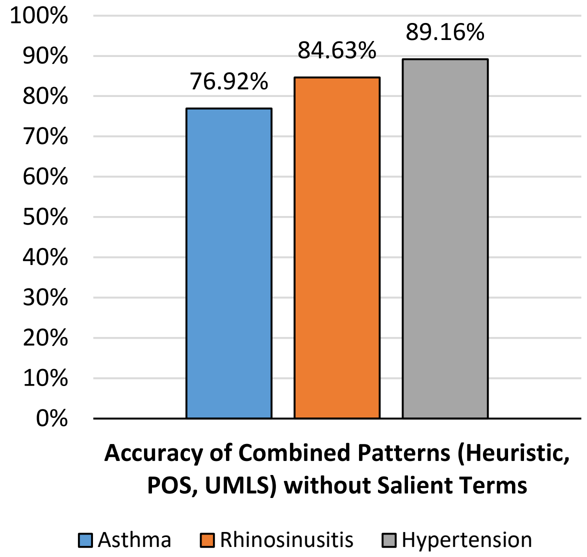
ML Assisted Patterns

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

1 Hussain, Musarrat, et al. "Recommendation statements identification in clinical practice guidelines using heuristic patterns." *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*. IEEE, 2018.

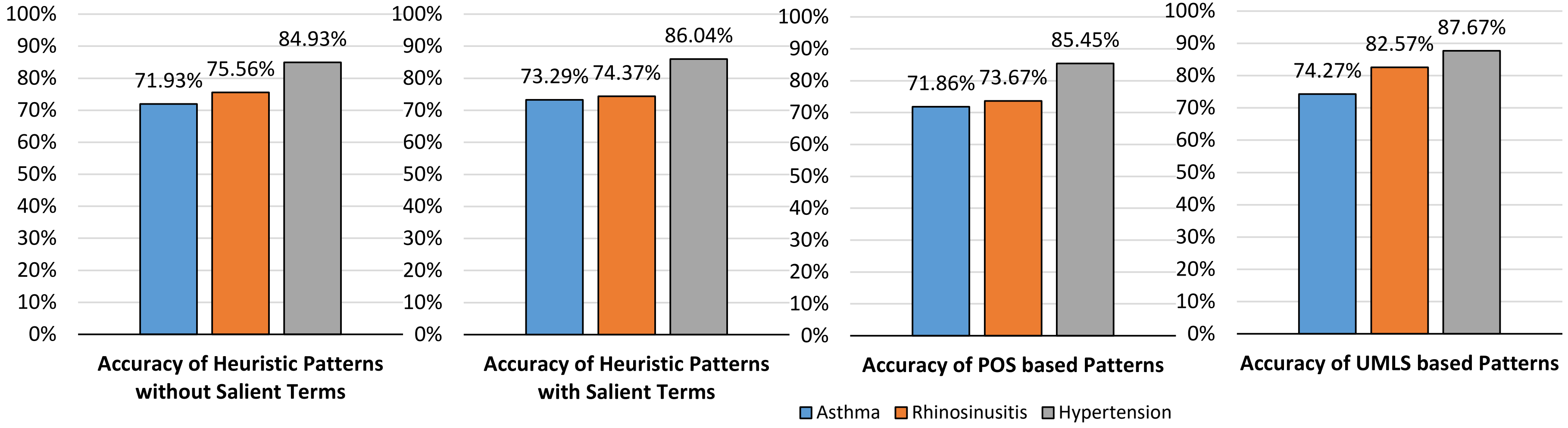
Experimental Results

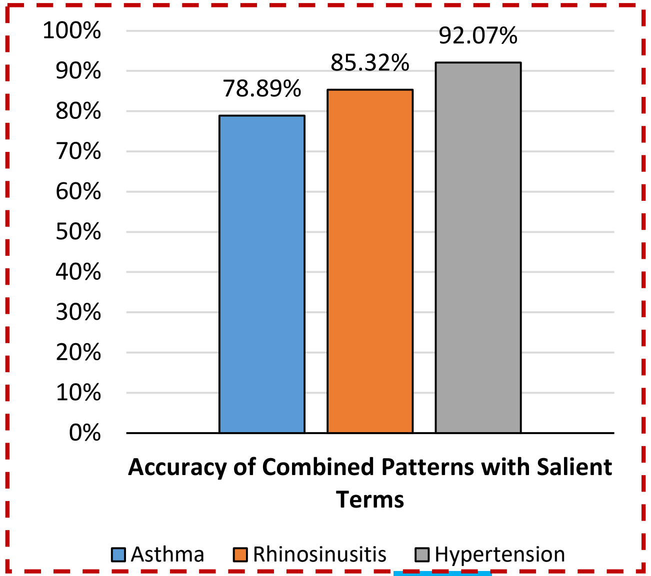
Solution 1: Machine Learning Assisted Pattern based Approach → **Results**



Experimental Setup

Guideline	Total Sentences	Recommendation Sentences	Non-Recommendation Sentences
Hypertension (Test set)	83	20(24.10%)	63(75.90%)
Rhinosinusitis	761	151(19.84%)	610(80.16%)
Asthma	171	53(30.99%)	118(69.01%)





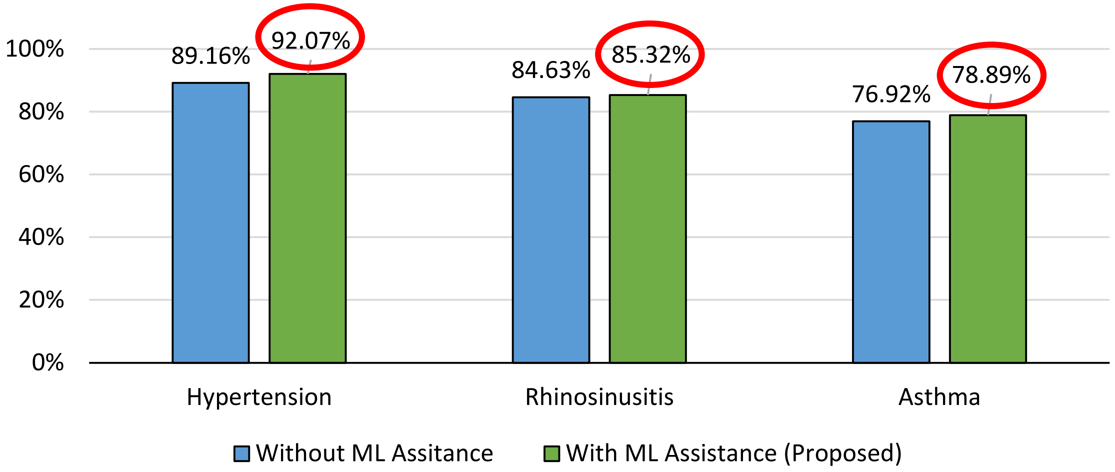
- The accuracy is calculated by **comparing the ground-truth** label of a sentence with the **predicted label**.

$$Accuracy = \frac{TP + TN}{P + N}$$

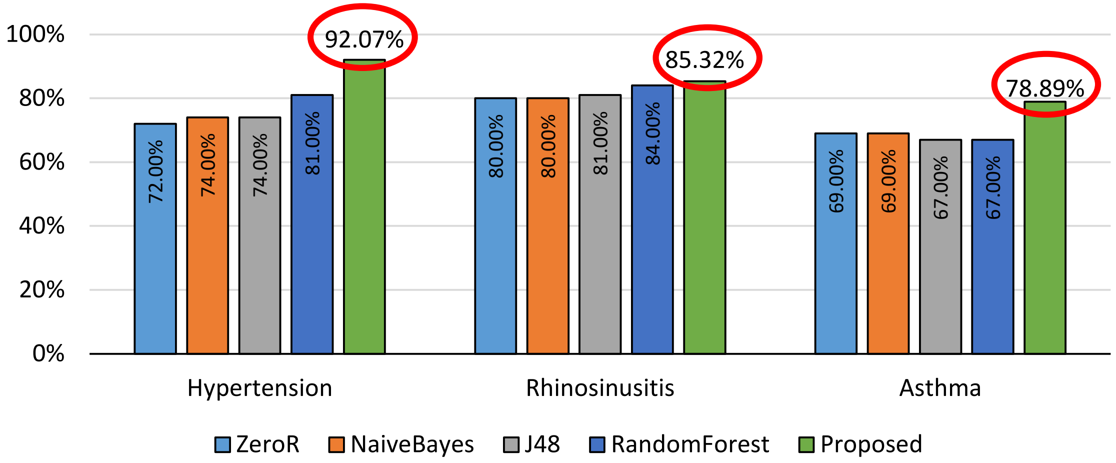
Hussain, Musarrat, et al. "Text Classification in Clinical Practice Guidelines Using Machine-Learning Assisted Pattern-Based Approach." *Applied Sciences* 11.8 (2021): 3296.

Experimental Results

Solution 1: Machine Learning Assisted Pattern based Approach → **Evaluation**



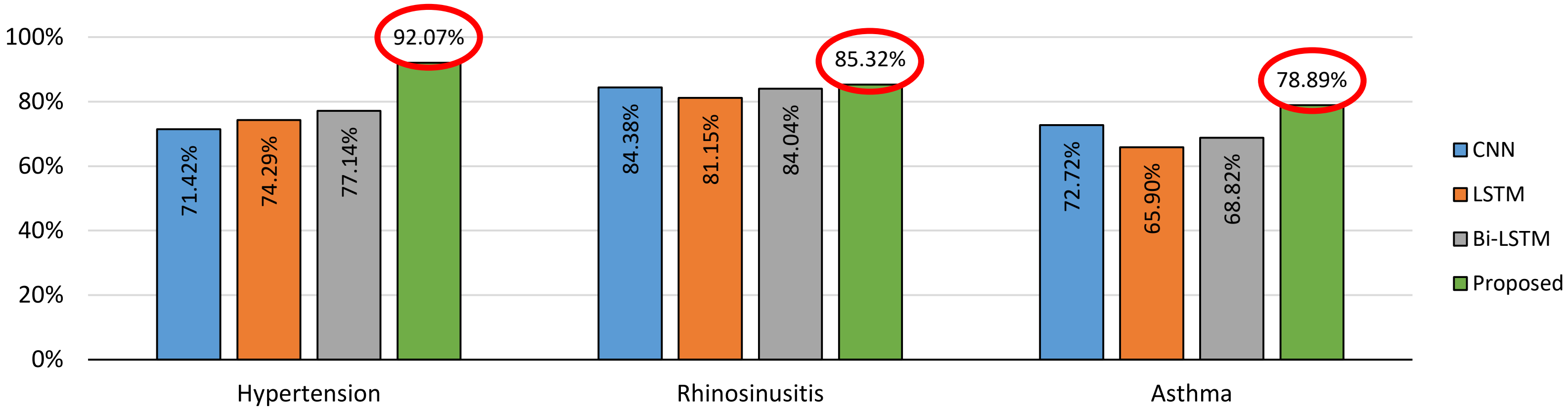
Sentence classification results without and with ML assistance



Classification results of proposed method vs traditional ML models

[50, 51]

- Embedding Size: 1776
- Activation: sigmoid
- Optimizer: adam
- Loss: binary_crossentropy
- Dropout: 0.5
- Metrics: accuracy
- Epochs: 100
- Batch Size: 10



Classification results of proposed method vs Deep Learning models

https://github.com/dipanjanS/nlp_workshop_dhs18/tree/master/Unit%2012%20-%20Project%209%20-%20Sentiment%20Analysis%20-%20Supervised%20Learning

Experimental Results

Solution 1: Machine Learning Assisted Pattern based Approach → **Evaluation with Large Dataset**

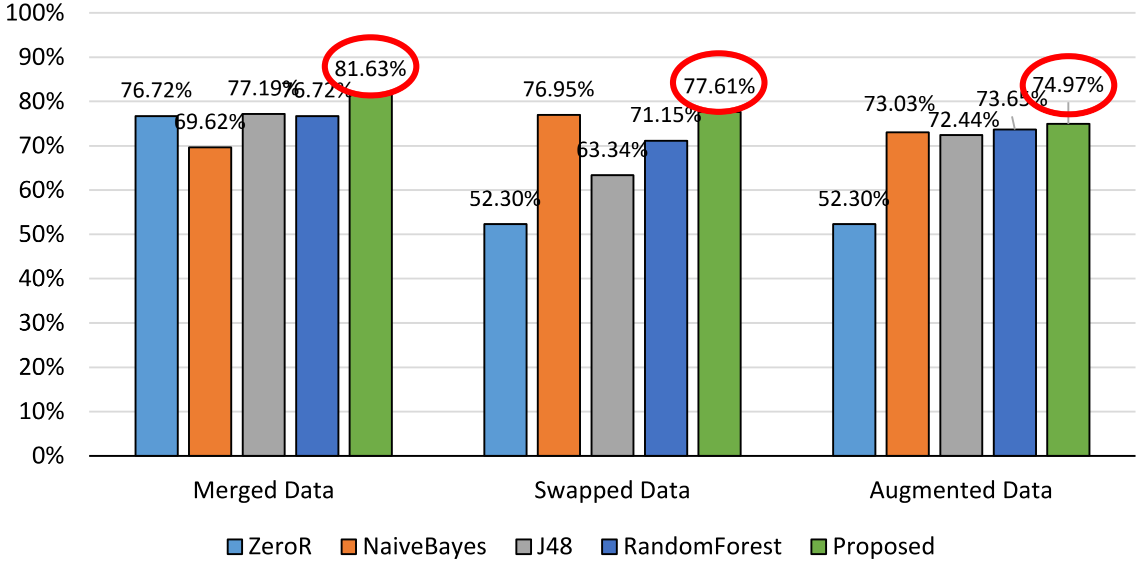
Experimental Setup

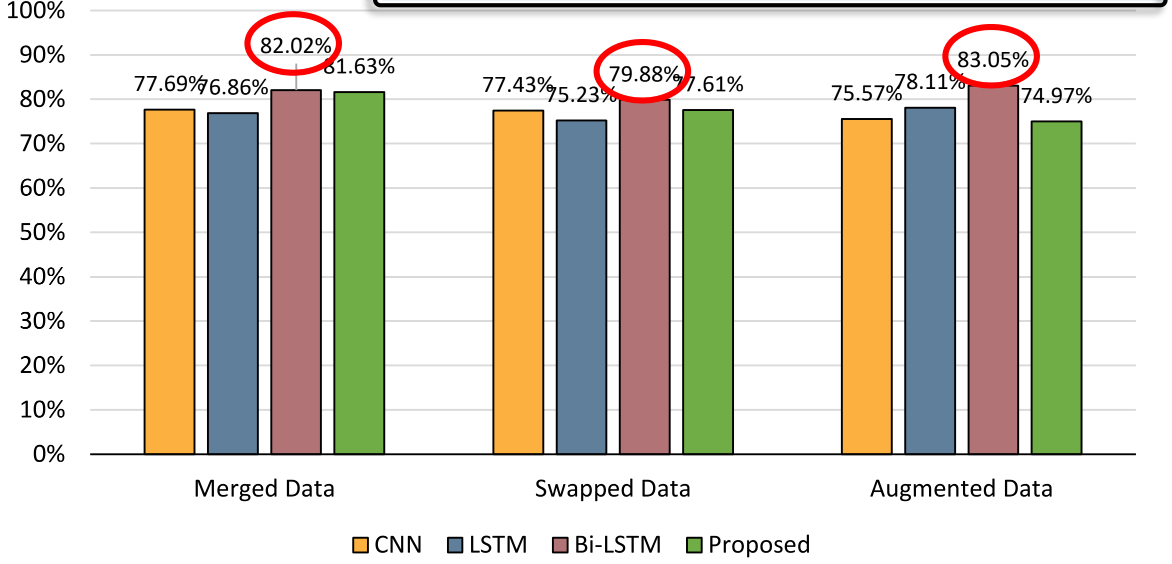
Method	Total Sentences	Recommendation Sentences	Non-Recommendation Sentences
Merged Data	1210	282(23.31%)	928(76.69%)
Swapped Data	1775	846(47.66%)	929(52.34%)
Augmented Data	1775	846(47.66%)	929(52.34%)

Merged all three datasets into a single comparatively large and an imbalanced dataset

Duplicate the number of RS sentences, and swap their token positions

Duplicate the number of RS sentences, and replace their tokens with synonyms



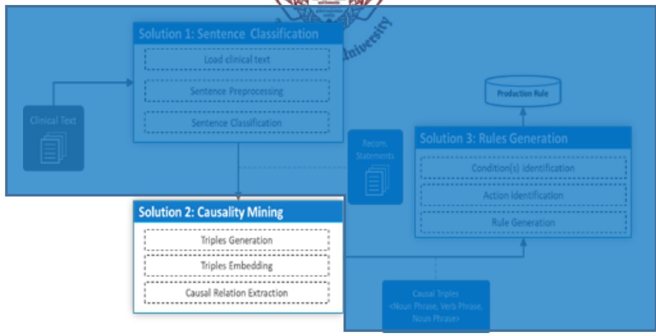


Result comparison with deep learning models on large and balanced datasets

Hussain, Musarrat, et al. "Text Classification in Clinical Practice Guidelines Using Machine-Learning Assisted Pattern-Based Approach." *Applied Sciences* 11.8 (2021): 3296.

Experimental Results

Solution 2: Causality Mining → Experimental Setup



1 Training Phase: Dataset Details

SemEval Training
Total Sentences = 8000
Causal Sentences = 1003
Non Causal Sentences = 6997

→

Extracted Triples
Initial Triples = 1071
Expanded Triples = 1246975

2 Threshold Selection: Dataset Details

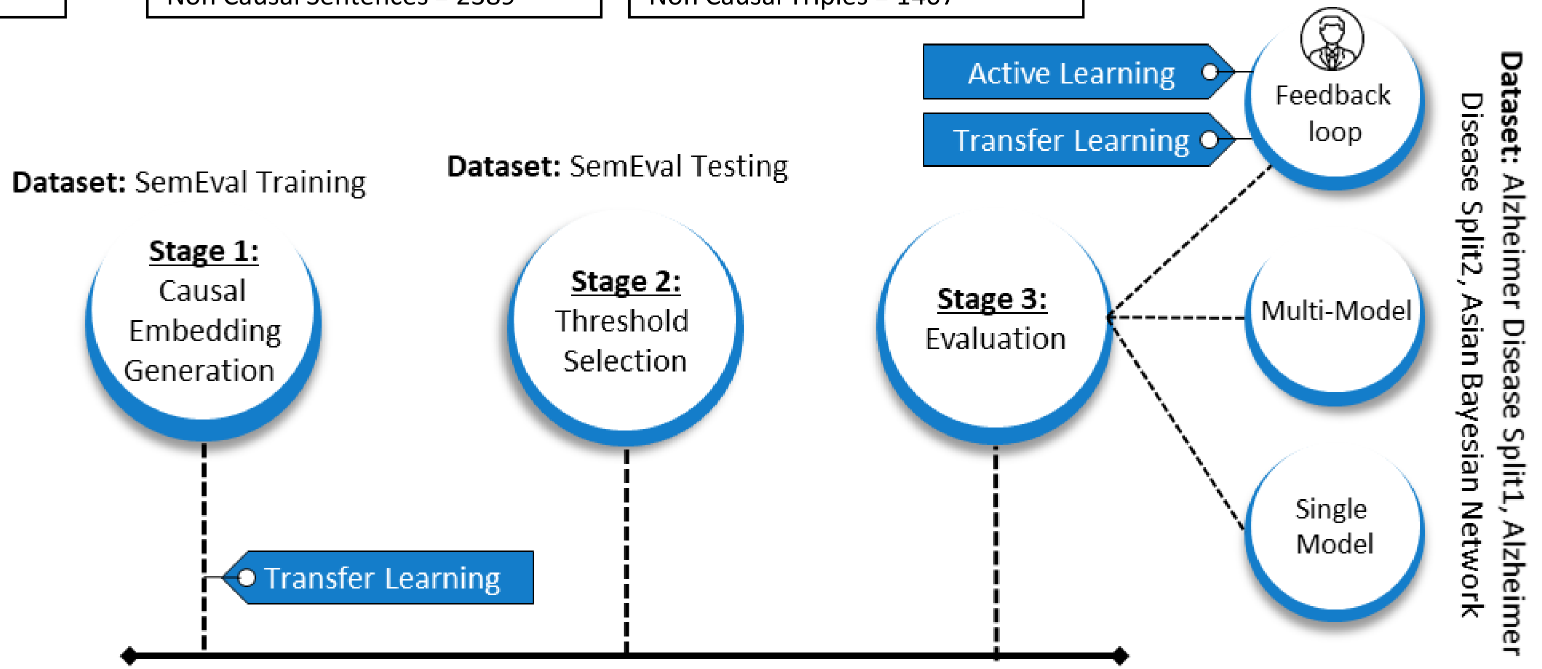
SemEval Testing
Total Sentences = 2717
Causal Sentences = 328
Non Causal Sentences = 2389

→

Extracted Triples
Total Triples = 1659
Causal Triples = 252
Non Causal Triples = 1407

3 Testing Phase: Dataset Details

Asian Bayesian Network	Alzheimer's Disease
Total Sentences = 500 Causal Sentences = 316 Non Causal Sentences = 184	Total Sentences = 2500 Causal Sentences = 1228 Non Causal Sentences = 1272
Total Triples = 86 Causal Triples = 47 Non Causal Triples = 39	Total Triples = 864 Causal Triples = 523 Non Causal Triples = 341



Experimental Setup

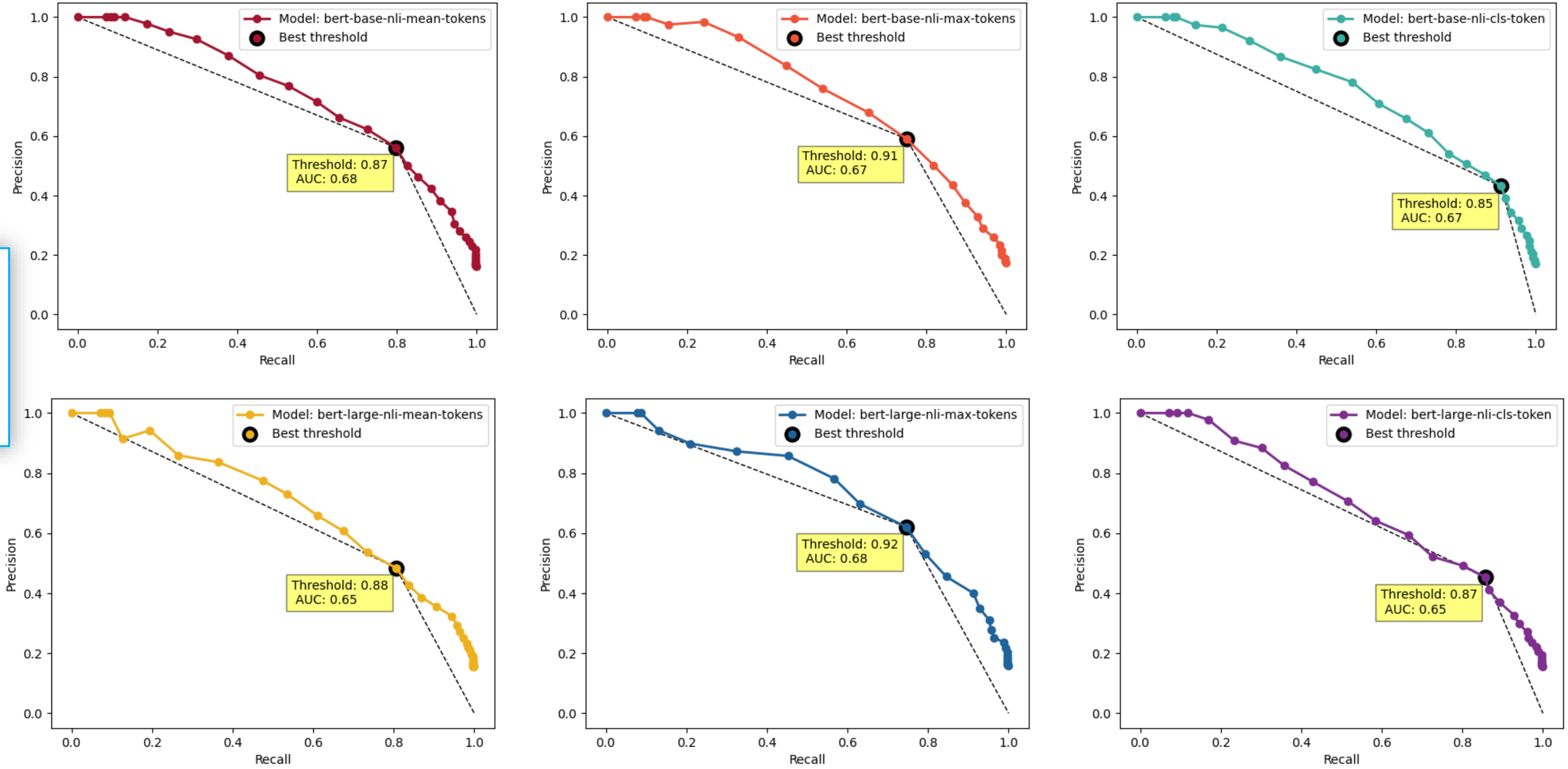
Hussain, Musarrat, et al. "A practical approach towards causality mining in clinical text using active transfer learning" Journal of Biomedical Informatics (2021): 103932.

Experimental Results

Solution 2: Causality Mining → Threshold Selection

PRC

Precision Recall Curve based Threshold Selection for Six BERT Models



1 Hussain, Musarrat, et al. "A practical approach towards causality mining in clinical text using active transfer learning" Journal of Biomedical Informatics (2021): 103932.

Experimental Results

Solution 2: Causality Mining → Causal Classification Results

BERT Models Results

Models	Asia Bayesian Network Dataset				Risk Factors of Alzheimer's Disease Split 1			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
BERT nli-base-mean-tokens	48.84	61.54	17.02	26.67	44.21	**63.27**	23.22	33.97
BERT nli-large-mean-tokens	58.14	58.73	78.72	67.27	45.37	58.12	41.57	48.47
BERT nli-base-max-tokens	44.19	47.37	19.14	27.27	44.44	61.54	26.97	37.50
BERT nli-large-max-tokens	53.49	60.00	44.68	51.22	44.21	59.70	29.96	39.90
BERT nli-base-cls-token	**62.79**	**64.15**	72.34	68.00	51.39	61.09	58.80	59.92
BERT nli-large-cls-token	59.30	59.38	**80.85**	**68.47**	**52.31**	61.34	**61.80**	**61.57**

Models	Risk Factors of Alzheimer's Disease Split 2			
	Accuracy	**Precision**	**Recall**	**F1**
BERT nli-base-mean-tokens	48.38	**68.97**	23.44	34.99
BERT nli-large-mean-tokens	54.17	64.65	50.00	56.39
BERT nli-base-max-tokens	49.31	66.67	28.91	40.33
BERT nli-large-max-tokens	48.61	61.97	34.38	44.22
BERT nli-base-cls-token	**57.41**	63.85	64.84	**64.34**
BERT nli-large-cls-token	54.86	60.78	**67.19**	63.82

1 Hussain, Musarrat, et al. "A practical approach towards causality mining in clinical text using active transfer learning" Journal of Biomedical Informatics (2021): 103932.

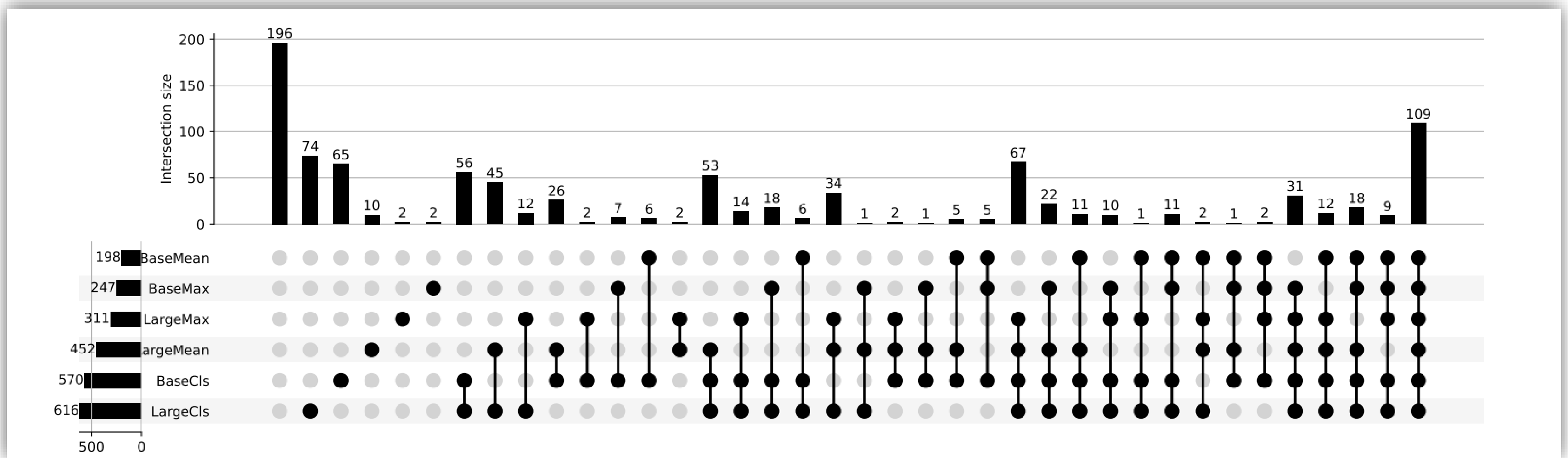
Experimental Results

Solution 2: Causality Mining → **Causal Classification Result Analysis**

UpSet Analysis[28]

- Analysis of the number of overlapping among various BERT models and the true positive causal triples.
- The number of actual causal triples uniquely identified by various BERT Models leads to a multiple model analysis

- Amongst the **950 candidate triples** of both the datasets, **754** have been classified as causal by one or more of the BERT NLI models.
- **True positive** classification for the six BERT 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.



Hussain, Musarrat, et al. "A practical approach towards causality mining in clinical text using active transfer learning" Journal of Biomedical Informatics (2021): 103932.

UpSet: http://vcg.github.io/upset/?dataset=10

Experimental Results

Solution 2: Causality Mining → **Multi-Model Causal Classification Results**

Multi-Model Results

Datasets	Accuracy	Precision	Recall	F1
AD1	56.25	61.40	78.65	68.97
AD2	56.25	59.78	80.08	68.45
Asia	60.47	59.42	87.23	70.69

Multi-Model with Feedback Loop (Active Learning)

Iteration	Dataset	Dataset Evaluation				Expert Evaluation	
		Accuracy	Precision	Recall	F1	Triples Added to CTTM	Triples 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**)	368	49
3	Asia	61.63%(↑**1.16**)	60.00%(↑**0.58**)	89.36%(↑**2.13**)	71.79%(↑**1.1**)	58	12

- **Iteration** represents an **execution** (triples classification by applying the **proposed causality mining** methodology including expert feedback) for an **unseen dataset.**
- At each iteration, our trained model **CTTM get evolved** as experts **verified true positive triples** are added and **falsely positive** and similar triples **are removed** from CTTM.

Discussion on Results

- Expert Evaluation is used to **verify** the system-identified causal triples.
- The triples where experts **agreed** with the system decision are converted into embedding vectors and **added** to **CTTM** to **broaden** its **scope** for subsequent iterations.
- While in a case where experts **negate** the **system decision**, the triples are added to **Block List**.
- The **Block List** is used to **remove** the incorrect triples from CTTM to **restrict inaccurate** triples prediction.

Experimental Results

Solution 2: Causality Mining → Results Evaluation

Multi-Model with Feedback Loop

Dataset	Ning's Method Evaluation				Proposed Method Evaluation			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**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**

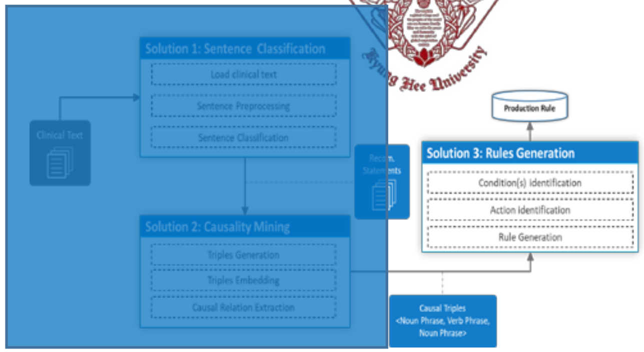
Discussion on Results

- We observed that **Ning's methodology** classifies **all triples as causal,** achieving a recall rate of 100%. However, the accuracy, precision, and F1 scores are decreasing by comparatively large margins.
- Therefore, we 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.
- Our methodology is able to **improve upon** its results across iterations.

An, Ning, et al. "Extracting causal relations from the literature with word vector mapping." Computers in biology and medicine 115 (2019): 103524.

Experimental Results

Solution 3: Rules Generation → **Results**



Experimental Setup

Guideline	Total Sentences	Recommendation Sentences	Non-Recommendation Sentences
Hypertension	278	78(28.06%)	200(71.94%)

Solution 3

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

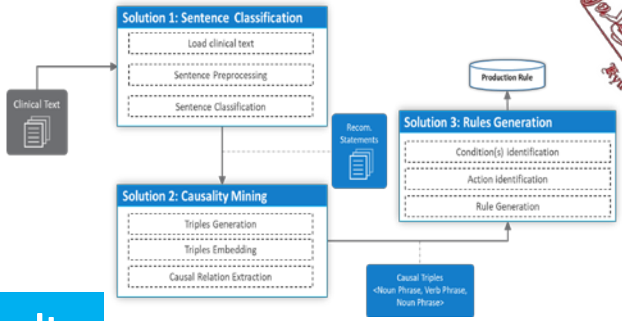
Manually annotated concepts with their values as ground truth dataset

Discussion on Results

- The proposed solution accurately identified concepts' values in cases where the values are located nearby the target concepts.
- Among the six missed values, five are located for-away from the concepts. Such as, "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."

Experimental Results

End to End Methodology → **Results**



- **Guideline 1:** Classification and Diagnoses of Diabetes: Standards of Medical Care in Diabetes – 2020[38]
(ADA)

- **Guideline 2:** Type 2 diabetes in adults: management [39]
(NICE)

- **Guideline 3:** Management of diabetes: A national clinical guideline[40]
(SIGN)

Experimental Setup and Results

Process	Guideline1	Guideline2	Guideline3	Total
Sentences	367	1805	279	2451
Extracted Triples	1731	1142	10226	13099
Unique Triples	1602	948	8872	11422
Medical Triples	1267	831	7765	9863
Casual Triples	541	320	7765	3215
Extracted Rules	**29**	**7**	**13**	**49**

Triples to Rule Example

S.No	Triple	Rule
1	<HbA1c, is, Diabetes>	IF HbA1c >= 6.5 THEN Diabetes
2	<FPG, be, Diabetes>	IF FPG >= 126 THEN Diabetes
3	<Greater RPG, Leads to , diabetes	IF RPG >= 200 THEN Diabetes
4	<FPG, be, Prediabetes>	IF FPG 100-125 THEN Prediabetes
5	<HbA1c, is, Prediabetes>	IF HbA1c 5.7-6.4 THEN Prediabetes

1

Hussain, Musarrat, et al. "Intelligent knowledge consolidation: From data to wisdom." Knowledge-Based Systems 234 (2021): 107578.

Experimental Results

Extracted Rules → **Evaluation**

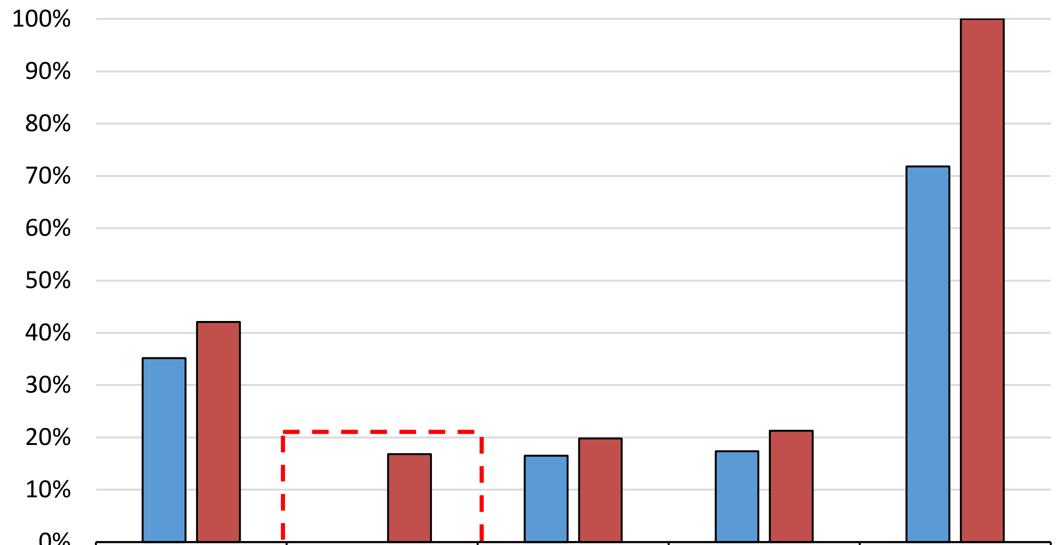
THE CATHOLIC UNIVERSITY OF KOREA
SEOUL ST. MARY'S HOSPITAL

Dataset Detail

Total Instances	Class Labels	Instance Ratio	Total Features
302	4 (No DM, Recheck, Pre DM, DM)	63:49:60:130	12

Current and Previous Values of

FPG, HbA1c, OGTT, PPG, Sign and Symptoms and RPG



	DM	Recheck	Prediabetes	No Diabetes	Overall
Extracted Rule Decision	35.15%	0%	16.50%	17.33%	71.79%
Expert Decision	42.08%	16.83%	19.80%	21.29%	100%

Extracted Rule Decision   Expert Decision

Result Discussion

- Expert Provided 16.83% times **"Recheck"** as a final decision, while guidelines and the extracted rules only recommend one of the three decisions (DM, Prediabetes, and No Diabetes).

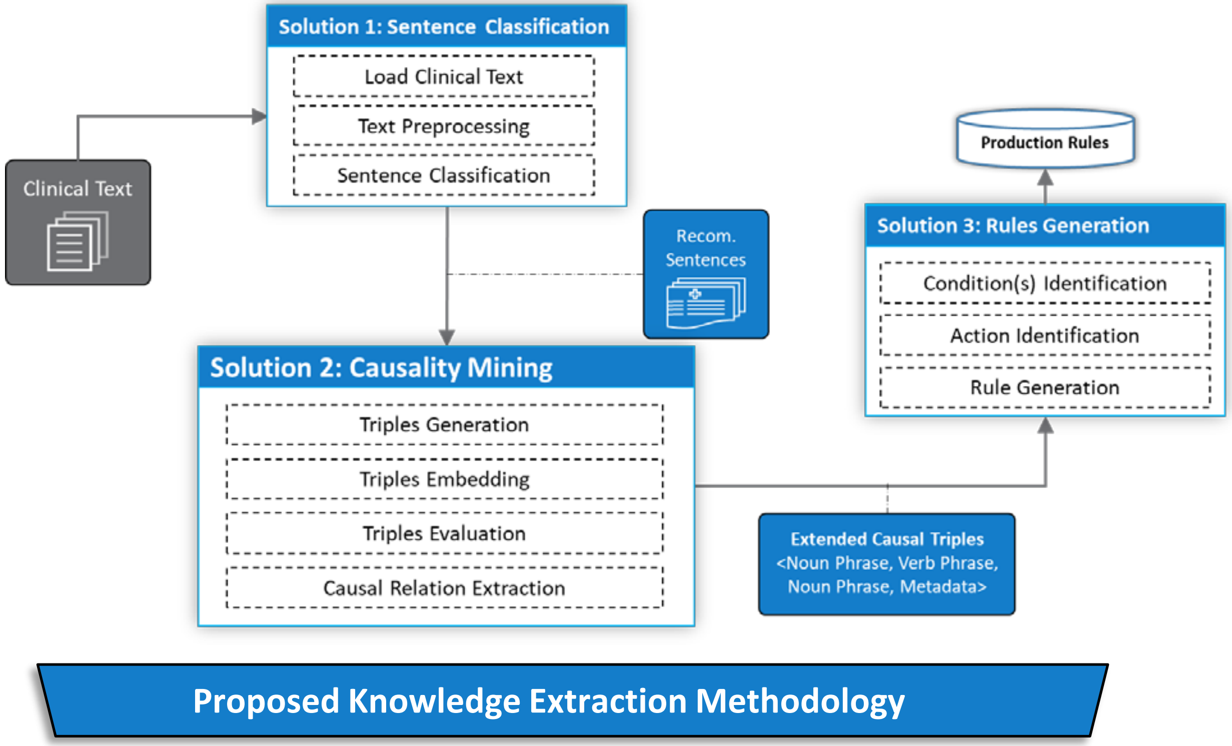
Abbreviations

FPG: Fasting Plasma Glucose
HbA1c: Hemoglobin A1c
OGTT: Oral Glucose Tolerance test
PPG: Post prandial Glucose
RPG: Random Plasma Glucose
DM: Diabetes Mellitus

1

Hussain, Musarrat, et al. "Intelligent knowledge consolidation: From data to wisdom." Knowledge-Based Systems 234 (2021): 107578.

Potential Applications



- Assist clinical experts in coping with the surfeit of unstructured textual information [1,2].
- Enables the use of clinical text for clinical decision support and quality improvement [2,3].
- Support Patients heath summary generation from associated unstructured documents [2]
- Provided the foundations for an accurate and evolvable knowledge-base, that can greatly enhance decision making in the healthcare domain [4,5].

[1] Neustein, Amy, et al. "Application of text mining to biomedical knowledge extraction: analyzing clinical narratives and medical literature." Text Mining of Web-based Medical Content (2014): 3-32.
[2] Kreimeyer, Kory, et al. "Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review." Journal of biomedical informatics 73 (2017): 14-29.
[3] Wang, Yanshan, et al. "Clinical information extraction applications: a literature review." Journal of biomedical informatics 77 (2018): 34-49.
[4] Hussain, Musarrat, et al. "Intelligent knowledge consolidation: From data to wisdom." Knowledge-Based Systems 234 (2021): 107578.
[5] Hussain, Musarrat, et al. "A practical approach towards causality mining in clinical text using active transfer learning." Journal of Biomedical Informatics 123 (2021): 103932.

Contributions & Uniqueness

Contribution

- Machine learning assisted pattern extraction and Automatic pattern extraction algorithms for clinical sentence classification.
- BERT based Multi-Model Active Transfer learning approach for causality mining.
- Causal triples to production rules conversion

Uniqueness

- Proposed an end-to-end methodology for knowledge extraction from clinical text produces transparent knowledge which can be used by automated systems for assisting clinical decisions as well as by human experts for quality service provisions.

Conclusions and Future Works

Sentence classification

- Proposed an automatic and machine learning assisted pattern extraction methodology for sentence classification

Causality mining

- Proposed an active transfer learning based approach for causality mining which achieved reasonable performance and increase the performance over iterations

Rule Generation

- The entities of each causal triple are evaluated for condition and corresponding actions.
- The appropriate value of each condition is identified via NLP Parser and UMLS dictionary

Future Works

- The presented clinical knowledge extraction pipeline can be further enhanced by replacing individual modules with other state-of-the-art methods.
- The acquired knowledge can be represented in more feature-rich models such as Knowledge Graph.

Publications

Patents (4)

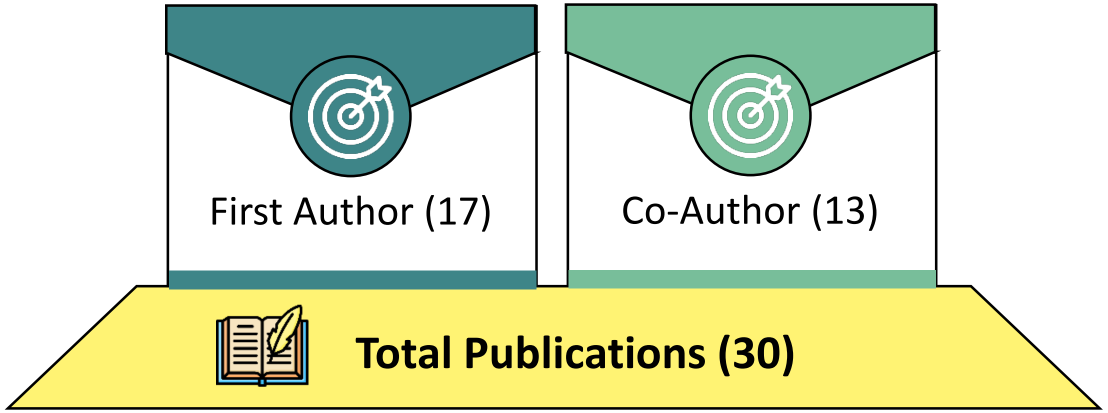
- International (2):
  - Registered → (1), Applied → (1)
- Domestic (2):
  - Registered → (1), Applied → (1)

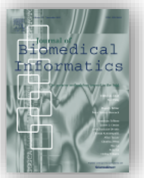
SCI/E Journals (13)

- First Author (5)
  - International → (3), Domestic → (2)
- Co-Author (8)
  - International → (6), Domestic → (2)

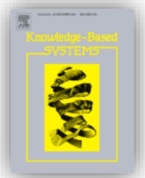
Conference (13)

- First Author (8)
  - International → (5), Domestic → (3)
- Co-Author (5)
  - International → (5)

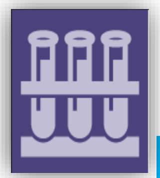




Journal of biomedical informatics IF: 6.317



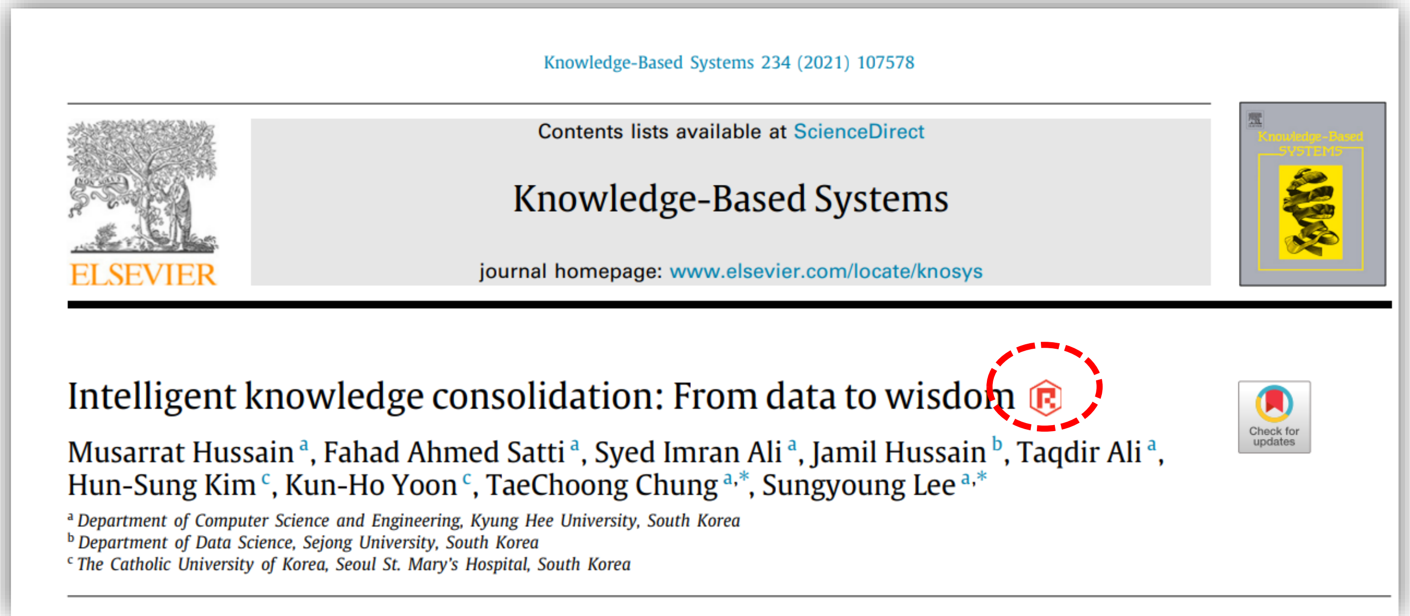
Knowledge-Based Systems IF: 8.038



Applied Sciences IF: 2.679

Achievements

Reproducibility Badge



Knowledge-Based Systems 234 (2021) 107578

ELSEVIER

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Intelligent knowledge consolidation: From data to wisdom

Check for updates

Musarrat Hussain [a], Fahad Ahmed Satti [a], Syed Imran Ali [a], Jamil Hussain [b], Taqdir Ali [a], Hun-Sung Kim [c], Kun-Ho Yoon [c], TaeChoong Chung [a,*], Sungyoung Lee [a,*]

[a] Department of Computer Science and Engineering, Kyung Hee University, South Korea
[b] Department of Data Science, Sejong University, South Korea
[c] The Catholic University of Korea, Seoul St. Mary's Hospital, South Korea

The code (and data) in this article has been certified as Reproducible by **Code Ocean**: (https:codeocean.com/). More information on the Reproducibility Badge Initiative is available https://www.elsevier.com/physical-sciences-andengineering/computer-science/journals

2016 The International Symposium on Perception, Action, and Cognitive Systems (PACS2016), October 27-28, 2016, aT center, Seoul, Korea

KIISE 16-1317

BEST PAPER AWARD
to

Musarrat Hussain, Maqbool Hussain and Sungyoung Lee(Kyung Hee University)

Paper Title: Smart CDSS: Knowledge Acquisition and Modeling using Mind Maps and Decision Trees

General Co-Chairs

Choong Seon Hong
Klaus Mainzer

PACS2016 is organized by Korean Institute of Information Scientists and Engineers (KIISE).

The international Symposium of Perception, Action, and Cognitive Systems PACS – 2016

References

1 Smith, Margaret, et al. "From code to bedside: implementing artificial intelligence using quality improvement methods." Journal of General Internal Medicine 36.4 (2021): 1061-1066.

2 https://artificial-intelligence.healthcaretechoutlook.com/cxoinsights/unstructured-data-in-healthcare-nid-506.html

3 Müller, Oliver, et al. "Using text analytics to derive customer service management benefits from unstructured data." *MIS Quarterly Executive* 15.4 (2016): 243-258.

4 Percha, Bethany. "Modern Clinical Text Mining: A Guide and Review." Annual Review of Biomedical Data Science 4 (2021).

5 Hudaa, Syihaabul, et al. "Natural language processing utilization in healthcare." (2019).

6 Assale, Michela, et al. "The revival of the notes field: leveraging the unstructured content in electronic health records." Frontiers in medicine 6 (2019): 66.

7 Thompson, Paul, et al. "The BioLexicon: a large-scale terminological resource for biomedical text mining." BMC bioinformatics 12.1 (2011): 1-29.

8 Spasic, Irena, and Goran Nenadic. "Clinical text data in machine learning: systematic review." *JMIR medical informatics* 8.3 (2020): e17984.

9 Sheikhalishahi, Seyedmostafa, et al. "Natural language processing of clinical notes on chronic diseases: systematic review." JMIR medical informatics 7.2 (2019): e12239.

10 Chintalapudi, Nalini, et al. "Text mining with sentiment analysis on seafarers' medical documents." International Journal of Information Management Data Insights 1.1 (2021): 100005.

11 Wang, Yanshan, et al. "A clinical text classification paradigm using weak supervision and deep representation." BMC medical informatics and decision making 19.1 (2019): 1-13.

12 Yang, Jie, Soyeon Caren Han, and Josiah Poon. "A survey on extraction of causal relations from natural language text." *arXiv preprint arXiv:2101.06426* (2021).

13 Sun, Wencheng, et al. "Data processing and text mining technologies on electronic medical records: a review." Journal of healthcare engineering 2018 (2018).

14 Zheng, Si, et al. "Text mining for drug discovery." Bioinformatics and Drug Discovery (2019): 231-252.

15 Bui, Duy Duc An, and Qing Zeng-Treitler. "Learning regular expressions for clinical text classification." Journal of the American Medical Informatics Association 21.5 (2014): 850-857.

16 Wenzina, Reinhardt, and Katharina Kaiser. "Identifying condition-action sentences using a heuristic-based information extraction method." Process support and knowledge representation in health care. Springer, Cham, 2013. 26-38.

17 Hematialam, Hossein, and Wlodek W. Zadrozny. "Identifying Condition-action Statements in Medical Guidelines: Three Studies using Machine Learning and Domain Adaptation." (2021).

18 Alashri, Saud, et al. "Snowball: extracting causal chains from climate change text corpora." 2018 1st International Conference on Data Intelligence and Security (ICDIS). IEEE, 2018.

19 Doan, Son, et al. "Extracting health-related causality from twitter messages using natural language processing." *BMC medical informatics and decision making* 19.3 (2019): 71-77.

20 An, Ning, et al. "Extracting causal relations from the literature with word vector mapping." Computers in biology and medicine 115 (2019): 103524.

21 Murtaugh, Maureen A., et al. "Regular expression-based learning to extract bodyweight values from clinical notes." Journal of biomedical informatics 54 (2015): 186-190.

22 Zheng, Guineng, et al. "Opentag: Open attribute value extraction from product profiles." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.

23 Berger, Abi, and Lawrence L. Weed. "Opening the Black Box of Clinical Judgement." BMJ: British Medical Journal: International Edition 319.7220 (1999): 1279-1279.

24 Lauer, Andreas K., and Dariah A. Lauer. "The good doctor: more than medical knowledge & surgical skill." Annals of eye science 2 (2017)

25 https://drdollah.com/clinical-care-processes/

References

26 http://openminted.eu/sneak-preview-openminted-knowledge-base-text-data-mining/

27 Yang, Jie, Soyeon Caren Han, and Josiah Poon. "A survey on extraction of causal relations from natural language text." *arXiv preprint arXiv:2101.06426* (2021).

28 Xu, Huimin, et al. "Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

29 Lex, Alexander, et al. "UpSet: visualization of intersecting sets." IEEE transactions on visualization and computer graphics 20.12 (2014): 1983-1992.

30 Kim, Youngjun, et al. "A Hybrid Model for Family History Information Identification and Relation Extraction: Development and Evaluation of an End-to-End Information Extraction System." JMIR Medical Informatics 9.4 (2021): e22797.

31 Zhan, Xianghao, et al. "Structuring clinical text with AI: Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases." Patterns 2.7 (2021): 100289.

32 Fu, Sunyang, et al. "Clinical concept extraction: a methodology review." Journal of Biomedical Informatics (2020): 103526.

33 Zhong, Ning, Yuefeng Li, and Sheng-Tang Wu. "Effective pattern discovery for text mining." IEEE transactions on knowledge and data engineering 24.1 (2010): 30-44.

34 https://towardsdatascience.com/deep-learning-vs-classical-machine-learning-9a42c6d48aa

35 Yao, Liang, Chengsheng Mao, and Yuan Luo. "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks." BMC medical informatics and decision making 19.3 (2019): 31-39.

36 Liu, Jiandong, et al. "Data-driven regular expressions evolution for medical text classification using genetic programming." 2020 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2020.

37 Serban, Radu, et al. "Extraction and use of linguistic patterns for modelling medical guidelines." Artificial intelligence in medicine 39.2 (2007): 137-149.

38 Zaharia, Oana P., et al. "Role of patatin-like phospholipase domain–containing 3 gene for hepatic lipid content and insulin resistance in diabetes." Diabetes Care 43.9 (2020): 2161-2168.

39 NICE, Diabetes in adults, 2016, no. March, URL http://guidance.nice.org.uk/ QS6.

40 S. I. G. Network, Management of Diabetes-a National Clinical Guideline (No 116), NHS Quality Improvement Scotland, Edinburgh, 2010.

41 Gallagher, Morris, et al. "The nominal group technique: a research tool for general practice?." Family practice 10.1 (1993): 76-81.

42 Morid, Mohammad Amin, et al. "Classification of clinically useful sentences in clinical evidence resources." Journal of biomedical informatics 60 (2016): 14-22.

43 VanDam, Courtland, et al. "Detecting clinically related content in online patient posts." Journal of biomedical informatics 75 (2017): 96-106.

44 Pawar, Sachin, et al. "Knowledge-based Extraction of Cause-Effect Relations from Biomedical Text." arXiv preprint arXiv:2103.06078 (2021).

45 De Silva, Tharini N., et al. "Causal relation identification using convolutional neural networks and knowledge based features." International Journal of Computer and Systems Engineering 11.6 (2017): 696-701.

46 Redd, Douglas, et al. "Regular expression-based learning for METS value extraction." AMIA Summits on Translational Science Proceedings 2016 (2016): 213.

47 Cai, Tianrun, et al. "EXTraction of EMR numerical data: an efficient and generalizable tool to EXTEND clinical research." BMC medical informatics and decision making 19.1 (2019): 1-7.

48 Chapman, Wendy W., et al. "A simple algorithm for identifying negated findings and diseases in discharge summaries." Journal of biomedical informatics 34.5 (2001): 301-310.

49 Prabhakar, Sunil Kumar, and Dong-Ok Won. "Medical Text Classification Using Hybrid Deep Learning Models with Multihead Attention." Computational Intelligence and Neuroscience 2021 (2021).

50 Li, Qian, et al. "A Survey on Text Classification: From Traditional to Deep Learning." ACM Transactions on Intelligent Systems and Technology (TIST) 13.2 (2022): 1-41.

51 Jang, Beakcheol, et al. "Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism." Applied Sciences 10.17 (2020): 5841.

Thank you

Comments & Suggestions