BUILDING COMPUTATIONAL REPRESENTATIONS OF MEDICAL GUIDELINES USING LARGE LANGUAGE MODELS AND TRANSFER LEARNING

by

Seethalakshmi Gopalakrishnan

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2023

Approved by:

_____
Dr. Wlodek Zadrozny

_____
Dr. Victor Zitian Chen

_____
Dr. Wenwen Dou

_____
Dr. Yaorong Ge

_____
Dr. Xi (Sunshine) Niu

_____
Dr. Albert Park

ABSTRACT

SEETHALAKSHMI GOPALAKRISHNAN. Building Computational Representations of Medical Guidelines using Large Language Models And Transfer Learning . (Under the direction of DR. WLODEK ZADROZNY)

This dissertation explores the potential of natural language models, including large language models, to extract causal relations from medical texts, specifically from Clinical Practice Guidelines. The outcomes of causality extraction from Clinical Practice Guidelines on gestational diabetes are presented, marking a first in the field. We also release, the first of its kind, an annotated corpus of causal statements in the Clinical Practice Guidelines.

We address the challenge of classifying causal sentences with a small amount of annotated data at the inter-sentence level by treating it as a cross-domain transfer learning problem. Obtaining these classified sentences is the first step in extracting causality. Furthermore, we delve into the importance of modal verbs and the degree of influence from cause to effect. We show the capability of three models (BERT, DistilBERT, and BioBERT) to identify the degree of influence in the text.

Lastly, we tackle the challenge of sparse annotated data for the causality extraction from Clinical Practice Guidelines by, again, using transfer learning. We investigate the correlation between data similarity and the efficacy of transfer learning. We also investigate a zero-shot and few-shot approach to cross-domain transfer learning, and quantify the link between data similarity and success rates. With the cross-domain few-shot transfer learning, we achieve an F1-score of 0.81, which suggests transfer learning as a possible solution to address the limited availability of annotated data.

ACKNOWLEDGEMENTS

I want to express my heartfelt thanks to Dr. Wlodek Zadrozny, my dissertation advisor, for his constant support throughout my PhD journey. He always made himself available to meet with me and provide feedback whenever I needed help. I'm also deeply indebted to Dr. Luciana Garbayo for her assistance with data selection and her constructive feedback.

A heartfelt thank you goes to my committee members - Dr. Victor Zitian Chen, Dr. Wenwen Dou, Dr. Yaorong Ge, Dr. Xi (Sunshine) Niu, Dr. Albert Park - for accepting the responsibility of serving on my dissertation panel, as well as offering invaluable suggestions.

I also wish to thank the National Science Foundation (NSF) and the principal investigator of the NSF project (Award No. 2141124), Dr. Wenwen Dou, as part of this work was sponsored by the NSF.

Nikhil's contributions to data annotation and Sreekar's assistance with data preprocessing have been very helpful, and I am grateful for their work.

I would also like to express my gratitude to Dr. Jody Marshall, the CIS Ph.D. program coordinator, Dr. Mohamed Shehab, the Ph.D. Program Director, Dr. Erik Saule, the CS Track coordinator, and Dr. Min Shin, for their support during my PhD study.

Finally, heartfelt thanks to my family and friends for their unwavering support and prayers. Special appreciation goes to my husband, my daughter Yadvee, and my parents for their unwavering support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAFP American Academy of Family Physician

ACOG American College of Obstetrics & Gynecology

ACR American College of Radiology

ADA American Diabetes Association

BERT Bidirectional Encoder Representations from Transformers

BI-LSTM Bidirectional Long-Short Term Memory

BioDRB Biomedical Discourse Relation Bank

CEGS Centers of Excellence in Genomic Science

CHF Chronic Heart Failure

CNC Causal News Corpus

CNN Convolutional Neural Network

CPG Clinical Practice Guideline.

CRF Conditional Random Fields

CTD Comparative Toxicogenomics Database

DocRED Document level Relation Extraction Dataset

EE Event Extraction

ELMo Embeddings from Language Model

GPT-4 Generative Pre-trained Transformer 4

IoM Institute of Medicine

KDD Knowledge Discovery in Databases

KEPT Knowledge Enhanced Prompt Tuning

LLAMA Large Language Model Meta AI

LLM Large Language Model

LSTM Long Short-Term Memory

MeSH Medical Subject Headings

N2C2 National NLP Clinical Challenge

NER Named Entity Recognition

NICU Neonatal Intensive Care Unit

NLP Natural Language Processing

NN Neural Network

PCG Predictive Causal Graph

PDTB Penn Discourse Tree Bank

POS tag Part Of Speech tag

RE Regular Expression

RE Relation Extraction

RMSE Root Mean Square Error

RNN Recurrent Neural Network

RNTN Recursive Neural Tensor Network

SCANER Semi-automated CAusal Network Extraction from Raw text

SVM Support Vector Machine

TACRED Text Analysis Conference Relation Extraction Dataset

UMLS Unified Medical Language System

USPSTF United States Preventive Services Task Force

CHAPTER 1: INTRODUCTION

1.1 Overview

Humans use natural languages to communicate and exchange information while programming languages instruct computers on what actions to perform. However, compilers or interpreters cannot comprehend natural languages like English in the same way humans do. As a result, computational models are employed to gain a limited understanding. Natural Language Processing (NLP) is an area of research in computer science and artificial intelligence that involves the processing of natural language. It involves translating the natural language into data that a computer can use to learn about the world. Sometimes this learning can be used by a computer to generate natural language text [2]. As stated in [1], [3], the challenges in NLP involve speech recognition, natural-language understanding, and natural-language generation. It may involve tasks like text and speech processing, morphological analysis, syntactic analysis, lexical semantics, relational semantics, and discourse.

This dissertation concentrates on the task of information extraction, particularly the extraction of causality, which seeks to identify and categorize the semantic relationships within a text. Information extraction involves obtaining valuable information from unstructured text, as stated in [4]. The objective of this dissertation is to extract causal statements from Clinical Practice Guidelines. We also explore the transfer learning for the causality extraction tasks because of the limited availability of the annotated corpus for medical text.

Clinical Practice Guidelines are a set of guidelines that guide physicians in the decision-making process. Various societies provide numerous such guidelines, each created by individuals with diverse backgrounds. This can lead to inconsistencies

in the guidelines, as noted in our work [5, 6]. Recognizing these discrepancies is crucial for establishing a common practice. To identify such disagreements and for other applications like question answering [7, 8], it's necessary to extract pertinent information from the text. Such extracted information can be used to improve the performance of our previous similar work on question answering [9]. This extraction might involve basic classification [10] followed by extraction from the classified sentences. In our recent work [11], we create a text-to-knowledge graph that can be used to make financial and strategic decisions.

Recently, pre-trained language models like BERT [12], which dynamically adjusts the weightings between each part of the output and all elements of the input based on their connection (attention), have demonstrated remarkable effectiveness. Similarly, Large Language Models (LLMs) like GPT-4 [13], which are pre-trained on extensive data and later enhanced through reinforcement learning feedback from both humans and AI to ensure adherence with human principles and policy compliance, and LLAMA2 [14], a model trained on 2 trillion tokens and available in three different sizes (7B, 13B, and 70B), have made significant strides in numerous tasks, particularly in information extraction.

Despite the emergence of the most recent LLMs, BERT continues to be one of the top-performing models for various applications, including causality extraction. In this thesis, we utilize various versions of BERT (DistilBERT[15], SpanBERT[16], BioBERT[17]) and Large Language Models like GPT-4 and LLAMA2 for the task of causality extraction.

1.2 Motivation

This thesis principally explores the topics of causality extraction and transfer learning. This section outlines the inspiration for selecting these topics, their relevance, and the potential impacts they can make.

1.2.1 Causality extraction

This work focuses on retrieving the causal statements to retrieve causes, effects, actions, conditions, and signals from the clinical practice guidelines. It also focuses on extracting the degree of influence introduced in Section 5.3. An example of a degree of influence is given below.

Example 1.2.1 *If a patient smokes more than 20 cigarettes per day and more than five drinks a day have a hazard ratio of 8.5* [18, 19].

In medicine, the best explanations are causal. Causal reasoning is used in producing and evaluating the impact of medical guidelines. The mechanistic model is preferred in medicine (even if it is probabilistic). Automated analysis is necessary (for various applications like comparing differences in guidelines, diagnostic support, etc.) since there are currently over *37,000* of medical guidelines indexed on PubMed as "practice guidelines" and two orders of magnitude of articles that are used to produce the guidelines. *Most of them use causal statements.*

Possible impact opportunities for retrieving causal statements from medical guidelines are listed below:

- Medical guidelines have changed over time. We can compare the differences in the guidelines by extracting the causal relations.
- Extracting causal relations can provide additional diagnostic support based on less frequent cause-effect relationships.
- Evaluating the strength of evidence is important and could be, together with an explanation, an important contribution to medical informatics can be made.
- In the medical field, the decision to provide the treatment is based on the fact that medication will improve the patient's condition (relationship between medication and improvement or not). Retrieving such relationships from the

medical guidelines can be used to construct a knowledge graph which will be helpful for doctors.

- If there is a graph indicating the relationship between the disease-cause, disease symptom, treatment-result, etc. Such knowledge graphs can also be compared to find whether there are any inconsistencies in the guidelines [6].

- Causal relationships can also be used for many other applications such as question answering [7, 8, 20, 21], information retrieval [22], medical text mining [23, 24, 25]

1.2.2 Transfer learning

This work applies transfer learning between the datasets of two different domains. In Section 2.4.3 and Section 3.2.2, we summarize some of the available datasets for causality extraction. A lot of causality extraction datasets are available, but only a very few causality extraction datasets are available for medical text. Out of the available annotated causality extraction datasets for medical text, none of the datasets are annotated on Clinical Practice Guidelines. A success in transfer learning will help expand the corpus of the annotated data for medical text.

1.3 Problems Statement

Given the motivation in Section 1.2, we realize the significance of extracting causal statements from medical guidelines for various applications. Given the importance of retrieving the causal statements, this work aims:

- To automatically identify the causal sentences from medical guidelines (this includes inter-sentence causalities) and to automatically extract which part of the extracted text is a cause/effect. Best explanations in medical practice are usually based on causality. Therefore, extracting causal statements can enhance

medical decision-making. Given the abundance of medical guidelines from various sources, an automated process for guideline analysis is essential. There is no work that focuses on causality extraction from Clinical Practice Guidelines, even though they have causal statements. Although there is considerable research on event extraction [26, 27, 28] and adverse drug event identification [29, 30, 31], in general, there is limited work on extracting causal relationships from medical text [32, 33, 34].

- To automatically identify the degree of influence of cause to effect. Example 1.2.1 is an example of the degree of influence. The degree of influence gives the extent to which the cause influences an effect. If we have this information on the amount of influence, the medical professionals can make a decision easily based on the positive or negative impact. It's an open research problem which is not yet explored.

- To identify the relationship between the train, test data, and the success of transfer learning between organizational data and medical guidelines data. Drawing on the results, we propose a model for zero-shot and few-shot cross-domain transfer learning, which has the potential to broaden the corpus of annotated data.

There are some deep philosophical discussions in causalities, such as there are multiple philosophical views and no common annotation standards. Since our work is more data-driven, we ignore these limitations.

1.4 Dissertation Contributions

This dissertation expands our prior publications related to the medical text[5, 6, 9] and causality extraction [11, 35], by presenting the results of causality extraction on an annotated corpus of medical text. This causality extraction can be incorporated with our prior work to improve its performance. The transfer learning approach proposed

in this dissertation can be used to expand the medical guidelines annotated corpus. The expanded corpus can subsequently contribute to bettering the performance of causality extraction.

- We created several models to automatically extract causality from medical text using transformers and LLMs (GPT-4 and LLAMA2). There is no work that focuses on extracting causalities from the Clinical Practice Guidelines (per Google Scholar search [1]). (Currently being prepared to be submitted to ACM Transactions on Computational Biology and Bioinformatics).

  - As part of the causality extraction, we created an annotated corpus that focuses on cause/effect relationships within Clinical Practice Guidelines for gestational diabetes. The data can be accessed at `https://github.com/gseetha04/Gestational-diabetes-annotated`. As per our research on Google Scholar[2], there are currently no other annotated datasets specifically designed for the task of causality extraction from Clinical Practice Guidelines.

- We developed a cross-domain inter-sentence causal sentence classification between the Organizational and medical guidelines data. According to a search on Google Scholar [3], no existing model employs cross-domain causal sentence classification at the inter-sentence level. In medical texts, relationships can often straddle multiple sentences. There are methods to extract causality from a document; however, very few strategies classify inter-sentence level causalities in sequential sentences prior to their extraction – particularly none for Clinical Practice Guidelines (CPGs). The importance of extracting causalities between consecutive sentences is underlined by the fact that many relationships might

[1]Search query: "causality extraction" from "clinical practice guidelines"
[2]Search query: causality extraction from clinical practice guidelines(Oct 20, 2023)
[3]Search query: Cross-domain inter-sentence level causal sentence classification (Oct 20, 2023)

be encapsulated in a follow-up sentence. Extracting relations from the entire document may pose challenges, but focusing on consecutive sentences might lead to the capture of numerous relationships. (Currently being prepared to be submitted to a Journal).

- We identify the degree of influence from cause to effect. Identifying the extent to which a cause influences an effect in medical texts is crucial. Recognizing these varying degrees of influence can aid medical professionals in their decision-making process. At present, there is no research dedicated to determining the impact degree from cause to effect, specifically in medical guidelines, an issue that this thesis seeks to rectify. (Currently being prepared to be submitted to a Journal jointly with inter-sentence level causal sentence classification).

- Driven by the lack of annotated data for medical texts (with no annotated data available for CPGs), we developed a cross-domain zero-shot and few-shot transfer learning method for causality extraction using transformer-based Large Language Models (LLMs). Our findings demonstrate that few-shot cross-domain transfer learning enhances the performance of causality extraction compared to zero-shot cross-domain transfer learning.

  - We present a study aimed at exploring the correlation between data similarity and the level of success achieved through transfer learning. (Submitted to Elsevier NLP journal - Article under revision)

  - In preparation for transfer learning, we present the results of causality extraction on organizational data. Within this work's scope, we introduce a new text-to-knowledge graph pipeline specifically designed for organizational data, which transforms a text document into a knowledge graph. This solution addresses a key requirement in recent accounting research, where decision-makers heavily depend on cause-and-effect insights (such as

materiality) inherent in accounting reports to make financial and strategic decisions [36]. (Published in Information Journal 2023 and a patent application)

1.5 Dissertation Structure

The remainder of this dissertation is organized as follows: Chapter 2.1 introduces and explores the concept of information extraction and related work. This chapter also discusses causality extraction, which is one of the primary themes of this work. Additionally, it includes an overview of previous work concerning causality extraction and elucidates on accessible datasets for said extraction.

Chapter 3 defines Clinical Practice Guidelines and summarizes related works on information extraction from these guidelines. It reviews previous work on event extraction, Named Entity Recognition (NER), and causality extraction from medical texts. The chapter concludes with a summary of datasets available for information extraction from medical texts.

Chapter 4 presents the key terms associated with causality extraction and defines our causal unit (the relations we intend to extract from the text) using examples from medical texts. Moreover, this chapter also explains the annotation process.

In Chapter 5, an automated method for identifying cause and effect relationships from two datasets is divulged, utilizing models such as BERT and Large Language Models like LLAMA2, GPT-4. This chapter also tackles the identification of sentences possessing a demonstrable influence factor (cause-to-effect impact) and presents inter-sentence relation classification. The chapter concludes with the presentation of the causality extraction results using both the organizational and medical datasets.

Addressing the limited availability of annotated data for medical text, Chapter 6 proposes utilizing a transfer learning approach for causality extraction from text. Initially, we present our findings and results regarding the relationship between the training data, test data, and the efficiency of transfer learning for causality extraction.

We observe an almost 35% increase in the macro-average F1-score when the gap between the training and test distributions is small according to the K-L divergence, the best-performing predictor on this task. To further probe this theory, we present the results of a cross-domain zero-shot and few-shot transfer learning approach between financial and medical domain texts.

Chapter 7 emphasizes the dissertation's key contributions while summarizing its overall findings. This closing chapter also speculates on potential future directions for this research area.

CHAPTER 2: INFORMATION EXTRACTION

From Chapter 1, we can understand that this work mainly concentrates on information extraction from medical guidelines. This chapter is meant to introduce natural language processing concepts. Since there are many concepts related to natural language processing, we constrained it to explain the concepts related to information extraction, as causality extraction is the main focus of this work.

2.1 Natural Language Processing and related concepts

Natural Language Processing is an area of research in computer science and artificial intelligence that is concerned with the processing of natural languages like English etc. It involves translating the natural language into a computer-understandable form, which can be used by a computer to learn about the world, which sometimes can be used to generate natural language text reflecting that understanding [2].

According to [1], there are many stages in natural language processing that can be viewed as layers in a feed-forward neural network. Figure 2.1 shows several stages of natural language processing.

Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [37]. Text mining or knowledge discovery from text deals with the machine-supported analysis of the text. It uses techniques from information retrieval, information extraction, and natural language processing, then connects them with the algorithms and methods to KDD, data mining, machine learning, and statistics [38].

Text mining [39, 1] is the processing of input text. Usually, it involves parsing, along with some derived linguistic features, and the removal of some features and

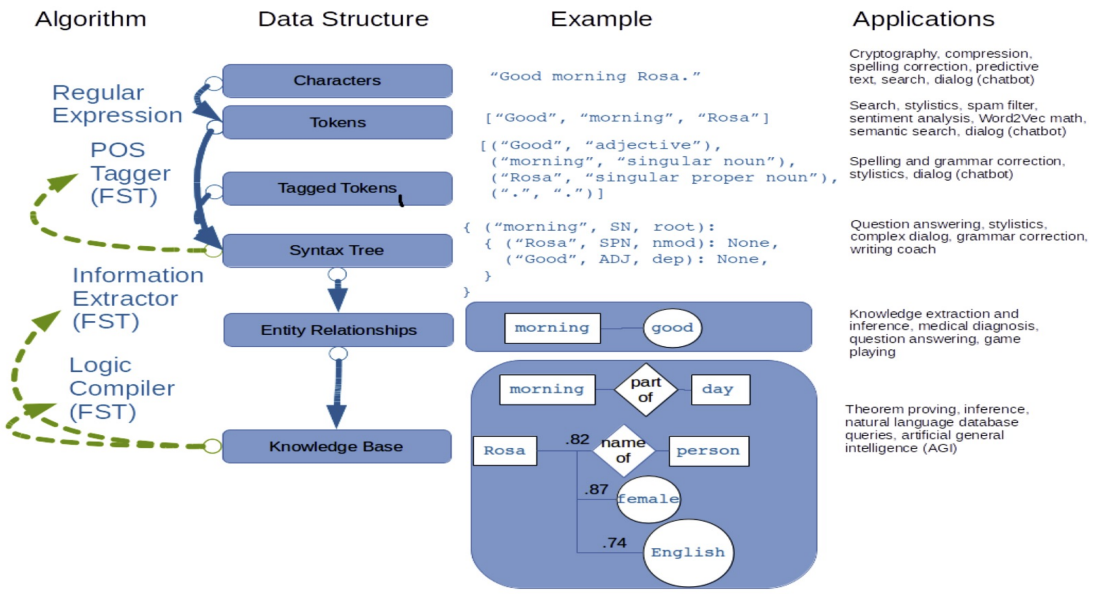


Figure 2.1: NLP Layers [1]. This figure shows layers of NLP that convert text into knowledge. It has been a popular architecture, which only recently is being replaced in parts by neural representations. From the knowledge base, we can use an information extractor to extract the entity relationships. For example, the causality extraction using BERT, discussed later, performs the relation extraction task, which can be used for various applications such as medical diagnosis.

intersection into the database. It also involves deriving patterns within the structured data and evaluating and interpreting the output. Text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities) are the typical text mining tasks.

Text analysis [40], [39] involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques, including link and association analysis, visualization, and predictive analytics. Essentially, the overarching goal is to turn text into data for analysis via natural language processing (NLP) applications, different algorithms, and analytical methods. An important phase of this process is the interpretation of the gathered information.

2.2 Information extraction

The process of turning the unstructured information embedded in texts into structured data is called information extraction. It is related to text simplification, which aims at creating machine-readable text. Typical information extraction tasks include relation extraction, event extraction, and template filling [4].

Relation extraction: The process of finding and classifying the semantic relations in the text is called relation extraction. It can be like a child-of-relation, part-whole, or geospatial relation.

Example 2.2.1 [$_{PERSON}$ *Bill Gates*] *announced that* [$_{PERSON}$ *John Smith*] *will be* [$_{PERSON}$ *the chief scientist*] *of* [$_{ORG}$ *Microsoft Corporation*] [41].

The above example gives the text and its relations. Bill Gates and John Smith are persons, and Microsoft is an Organization. This example gives generic relations.

For some of the categories of text, entity types, and the relations are available. For example, the Unified Medical Language System (UMLS) [42] categorizes the medical text into various categories, entity types, and relations. Tools like Metamap[43] are available for the medical text, which can be used to map the medical text to UMLS Metathesaurus, which is a database containing information about medical concepts, names, and their relations. The relations can be extracted in any of the following ways [44],

- Pattern-based relation extraction
- Supervised machine learning
- Semi-supervised relation extraction (bootstrapping or distant supervision)
- Unsupervised relation extraction

In this work, we will be concentrating on the supervised machine learning-based relation extraction task where the task is to use transformer-based deep learning models and Large Language Models to extract the cause/effect relations from medical text.

Apart from relation extraction, there are also other types of extraction strategies that are available. One of the common extraction is event extraction which is the process of extracting the events from the entities. In order to determine which events refer to the same event, an event coreference is needed. Temporal expressions are needed to figure out when an event in a text happened. An example showing the event mentions is given below.

Example 2.2.2 [$_{EVENT}$ *Citing*] *high fuel prices, United Airlines* [$_{EVENT}$ *said*] *Friday it has* [$_{EVENT}$ *increased*] *fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately* [$_{EVENT}$ *matched*] [$_{EVENT}$ *the move*], *spokesman Tim Wagner* [$_{EVENT}$ *said*]. *United, a unit of UAL Corp.,* [$_{EVENT}$ *said*] [$_{EVENT}$ *the increase*] *took effect Thursday and* [$_{EVENT}$ *applies*] *to most routes where it* [$_{EVENT}$ *competes*] *against discount carriers, such as Chicago to Dallas and Denver to San Francisco* [4].

2.3 Discussion of prior art on information extraction

This section summarizes some of the prior art on information extraction tasks, techniques, etc. The information extraction can be done using various techniques. A traditional way to do information extraction is to write rules and extract the information[45, 46], which is a rule-based information extraction. The rule-based approaches are slowly being replaced by machine learning techniques. It can be a supervised machine learning [47, 48] or unsupervised machine learning[49, 50]. Due to the availability of the labeled data for the supervised machine learning tasks, a semi-supervised machine learning technique [51, 52]leverages both the labeled and

the unlabeled data in order to reduce the annotation cost. In recent times, deep neural networks have been used for the information extraction task. Recent work on information extraction [53] uses a deep neural network to extract the syntactic and semantic elements from the data. It develops a corpus of manually annotated data and tests the model on it. They use a Bi-LSTM with Condition Random Fields (CRF), which obtained a precision and recall of 0.93 and 0.92, respectively. Zhang et al. [54] uses a Generative Transformer model (CGT) for extracting the information from text. This model does a triplet extraction and event extraction on text. The encoder of the CGT uses a BERT-like input representation along with word piece tokenization. In order to distinguish the encoder-decoder representation, a partial causal masking is used. It is tested on the New York Times (NYT) and WebNLG datasets. This model got an 89.1 F1 score on NYT and an 83.4 F1 score on the WebNLG dataset. A survey on the available web information extraction systems [55] summarizes the types of information extraction techniques available and tools used for the extraction. It compares the information extraction system based on three dimensions which are task difficulties, the techniques used, and the degree of automation used.

The above-mentioned works use some techniques to extract information from the unstructured text which is the main aim of this work.

2.4 Causality Extraction

In the previous section, information extraction was introduced, and some of the related work related to information extraction was summarized. The main topic of this thesis is causality extraction which is a type of information extraction. This section summarizes the methodologies related to causality extraction and some of the prior arts related to causality extraction. It also summarizes some of the available datasets that are related to causality extraction. A more detailed explanation of the causality extraction with examples along with the causal unit we define is explained in Chapter 4.

Information extraction plays an important role in natural language processing because of the large availability of unstructured text. The tasks of the information extraction [56] can be categorized into various categories, which include,

- Relation Extraction (RE)
- Named Entity Recognition (NER)
- Event Extraction (EE)

This work is concerned with the relation extraction task to extract the relationship between cause and effect in a text. The remainder of this chapter concentrates on causality extraction.

2.4.1 Three causality extraction methodologies and its related work

Based on the causality extraction techniques and the causality forms, we can classify the existing methods, which are summarized in this section. A recent survey [56] categorizes the existing causality extraction methodology into *knowledge-based, statistical machine learning-based, and deep learning-based methodologies*. A knowledge-based system can be pattern-based or rule-based. Sentence structure analysis, like lexico-semantic or syntactic analysis, can be used to explore pattern-based systems. The sentence structure can also be explored based on some procedures or heuristic algorithms. These types of systems are called rule-based systems. To identify the intra-sentence causality (i.e., explicit causality) [57, 58, 25] uses a pattern-based approach. Explicit causality is extracted by [59] by defining an algorithm based on the rule-based approach. Implicit causality is the process of extracting causal relations where the explicit signal/trigger defining the relation between the cause/effect will not be present. Various works on extracting implicit causality are available. To extract implicit causality [60, 61] uses a pattern-based approach. Inter-sentential causality extraction aims to extract the causalities present across the sentences. A pattern-based approach was used by [62] to extract the inter-sentential causality.

Statistical machine learning-based systems will use NLP tools like spaCy [63], Stanford CoreNLP [64], Stanza [65], etc., to create features for the machine learning algorithms. Rule-based approaches can be combined with machine learning techniques to extract the explicit causality [20], define the most common patterns in the data, and validate it with decision trees to extract the causal relations for a question-answering task. Implicit and explicit relations can also be extracted from various languages like Arabic [66], Thai[67] etc. Similarly, intra-sentence (causality within the sentences) and inter-sentence (causality across sentences) causality extractions can also be done for different languages like Japanese [68].

Neural Networks (NNs) will be the basic algorithm for deep learning-based approaches. When a neural network learns from multiple hidden layers, the neural network is said to be a deep neural network. Commonly used deep learning models include Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNNs), Gated Recurrent Units (GRU), Recurrent Neural Networks (RNNs), etc. In recent times, unsupervised pre-trained models like BERT significantly improved the performance of NLP tasks. Deep learning-based models are used by [69, 70, 71] for the causality extraction. A relatively recent work [72] uses bidirectional GRU with self-attention as the baseline and explores the application of using models such as BERT and ELMO for the causality extraction task. Many recent work on causality extraction uses models like BiLSTM to learn the entities and the relations between text[73, 70] to extract the implicit causality.

2.4.2 Causality extraction from text - Related work

Causality extraction is the task of automatically extracting the cause/effect relationships from the text. In this section, we will provide a summary of numerous studies related to causality extraction that are akin to our research. Nonetheless, none of the studies summarized here delve into the degree to which a cause influences an effect (see Chapter 4 for an in-depth explanation). Below are the distinctions

between our work and the work that has been summarized. [74], [75] extracts events from the text, but they don't extract cause and effect. A more similar work [70] extracts causalities from the given text using SemEval data. But they define a different causal unit than ours. They do not deal with scenarios like modals, inter-sentence causalities, and degree of influence.

The study by [70], which closely aligns with our work, directly extracts cause and effect from text without separately extracting candidate pairs and their relations. This is achieved through a neural causality extractor that uses BILSTM-CRF as its foundation. Their model, known as SCITE (Self-attentive BILSTM-CRF with Transferred Embeddings), expands the annotations of the SemEval 2010 task 8 dataset for their experiments. The findings indicate that SCITE, when combined with Flair embeddings [76], achieved an F1 score of 0.8455, surpassing the baseline scores.

Event causality extraction aims to extract the event relations from the text. A work on event extraction [74] focuses on identifying the causal relationship between pairs of event mentions, also known as Event Causality Identification (ECI). Initially, the input sequence is converted into an input-output pair (I, O). Following this, a transformer-based model, T5 [77], is applied to each transformed pair in the training data. The REINFORCE algorithm is used during the training of the T5 model to ensure label accuracy, which is used as the reward function. The model achieved an F1 score of 58.8 on EventStoryLine data and 56.5 on Causal-TimeBank. Balashankar et al. [75] propose an event extraction that seeks to uncover the hidden relationships between events mentioned in news streams by creating a Predictive Causal Graph (PCG). The PCG assesses how the appearance of one word can impact another word in future contexts. The dataset used for the experiments is a news dataset compiled from news articles. The root mean squared error (RMSE) is then calculated. When compared to the baseline model, the PCG derived from unigram (0.022), bigram (0.023), or both (0.020) exhibits a lower error.

The relationships in the text can also be represented in a tree format. Causality Treebank [78] that represents the data in a tree format in order to get the causal relations in a fine-grained form. The constructed dataset is tested using the Recursive Neural Tensor Network (RNTN) model on the corpus, which achieves a mean F1-score of 0.74 across all label types.

A recent work on causality extraction [79], Semi-automated CAusal Network Extraction from Raw text (SCANER), integrates both automated and manual methods to create a framework for generating a causal network from raw text. It is assessed on three different datasets encompassing political, food insecurity, and medical domains. SCANER achieves an F1-score of 0.82 on the political narrative dataset, 0.88 on the food insecurity text, and 0.89 on the medical text. Another similar work [80] extracts both the explicit and implicit causalities by representing the sentences as character, word, and contextual string layers. These representations are converted into a causality tree from which the labeled causalities are obtained. This multi-granulation representation is input into a neural network to extract causality. The model is evaluated using the SemEval 2010 Task 8 dataset.

In recent times, prompt tuning has been proposed to bridge the gap between pre-training and fine-tuning on many of the mainstream NLP tasks like text classification[81, 82], information extraction[83, 84] etc. A recent work [85], Knowledge Enhanced Prompt Tuning (KEPT) employs external knowledge sourced from knowledge bases (KBs) to fine-tune pre-trained language models through the design of an attention mechanism. The model achieves an F1-score of 0.58 on the EventStoryLine dataset and 0.54 on the Causal-TimeBank dataset.

2.4.3 Causality extraction from text - Datasets

This section summarizes some of the available datasets on causality extraction that are publicly available. This includes some of the datasets published as part of a competition/challenge. A summary of the available datasets is shown in Table 2.1.

The SemEval Task 8 [86] dataset is centered around the multi-faceted classification of semantic relationships between nominal pairs. This dataset comprises 10,717 annotated instances, with 8,000 samples designated for training and the remaining 2,717 for testing. The dataset defines nine relationships: Cause-Effect (CE), Instrument-Agency (IA), Product-Producer (PP), Content-Container (CC), Entity-Origin (EO), Entity-Destination (ED), Component-Whole (CW), Member-Collection (MC), and Message-Topic (MT). The initial data selection phase involved choosing 1,200 sentences for each relationship based on a pattern-oriented web search. In the subsequent phase, two annotators annotated the gathered candidates for each relationship. This dataset was developed as part of a challenge with the objective of extracting relationships from text.

The Penn Discourse Tree Bank (PDTB) [87] is a corpus featuring discourse-level annotated data. It includes annotated discourse relations and their arguments derived from the Wall Street Journal (WSJ) Corpus. The discourse relations in the PDTB are classified into two types: "Explicit" or "Implicit". This classification is based on the connectives. If a text lacks an explicit connective, it is inferred by the annotator, who then inserts an implicit connective. PDTB-2.0 has a total of 40,600 annotated relation tokens, comprising 18,459 explicit connectives, 16,224 implicit connectives, 624 AltLex relations, 5,210 EntRel relations, and 254 NoRel relations.

TACRED [88] is a supervised relation extraction dataset consisting of 119,474 examples. This dataset is derived from the TAC KBP challenge, utilizing query entities from 2009 to 2012 as training data, data from 2013 for development, and data from 2014 for testing. Each sentence in the dataset is annotated with the span of the subject and object entity, as well as the types of relations.

Document level Relation Extraction Dataset (DocRED) [89] is a dataset derived from Wikipedia and annotated with both named entities and relations. The annotation process involves generating distantly supervised annotations, followed by

annotating named entities and coreference information in Wikipedia articles. The annotated named entities are then linked to Wikidata items. Finally, relations and supporting evidence are labeled. The training data consists of 101,873 documents annotated in a weakly supervised setting and 3,053 documents annotated in a supervised setting. Both the development and test data include 1,000 documents annotated in both supervised and weakly supervised settings.

Automatic Content Extraction (ACE) [90] is a dataset created as part of a challenge. This corpus is produced in two genres: newswire and weblogs. Fifteen Chinese-English and fifteen Arabic-English translation document pairs were manually selected from the ET evaluation corpus. Out of the fifteen Chinese-English files, 786 Chinese entity mentions and 820 English entity mentions were identified. Among these, 772 mentions matched across the Chinese and English pairs, with 0.84 of them perfectly aligned. In the fourteen Arabic-English documents, there were 1,182 Arabic entity mentions and 1,521 English entity mentions. In total, there were 1,081 mention pairs across Arabic and English, with 0.74 of these pairs perfectly aligned.

The Inter-Sentence Relation Extraction Dataset [91] was developed using the resources utilized by [92]. This balanced dataset comprises 31,970 pairs of sentences with inter-sentence relation mentions involving 17 different relations. For the evaluation process, 100 pairs of sentences with explicit relation mentions for each of the relations were manually selected for the test set. The initial set of inter-sentence relations retrieved consisted of 87,514 sentence pairs. To maintain balance in the dataset, an equal number of sentence pairs were selected for each of the relations, resulting in a total of 31,970 pairs, out of which 14,861 are unique entities. Baseline models such as the Bag-of-Words model, LSTM, and BiLSTM were tested on the dataset, with the BiLSTM model achieving an F1 score of 0.72.

The 2023 Causal News Corpus (CNC) dataset [93] was released in conjunction with the Challenges and Applications of Automated Extraction of Socio-political

Events from Text (CASE 2023) workshop. The second version of this dataset includes annotations for Cause, Effect, and Signals. The dataset comprises a total of 1,981 sentences and 2,754 causal relations.

The Organizational dataset [11] utilizes the 2020 SEC 10-K documents of 65 S&P Financial Companies. Causal insights from the documents were annotated based on a pre-established dictionary of causal trigger words or cue phrases. A total of 2,234 sentences were identified and manually annotated as having a causal nature. For each of these identified causal sentences, the cause/effect relationship was indicated using tags. During the manual annotation process, effects were marked with the <outcome> tag.

The summary of some of the available datasets for relation extraction is summarized in Table 2.1.

Datasets [86], [96], and similar datasets published as part of the SemEval competition are the most commonly used datasets for the entity relationship extraction from the text. The causal unit defined in this task differs from what is used in this work. PDTB-2.0 [87] is the most commonly used discourse treebank dataset. Here the annotation for both implicit and explicit connectives is done at the discourse level. This dataset is closely related to the proposed relation extraction, but their causal unit consists of various senses. We are adapting some of the causal unit definitions from this corpus's guidelines. TACRED [88] is also an entity relation extraction dataset that is less related to our annotation. They have the person or the organization entities. DocRED [89], Inter-Sentence Relation Extraction Dataset [91] are entity relation datasets at a document and inter-sentence levels, respectively. Both of them are less related to our annotation because they use different causal units and annotation guidelines. ACE [94] is an entity, relation, and event extraction dataset. All the dataset listed in Table 2.1 is related to the task described in this work. However, they use a different type of causal unit or concentrate on entity relation or event

extraction. PDTB 2.0 [87] is more closely related to the causal unit discussed in this work, but it also includes many more sense types and relations.

From the listed datasets, we can understand that a lot of the datasets are available. The dataset that is closely related can be used for transfer learning tasks which will reduce the cost of the annotation. In our transfer learning experiments (explained in Chapter 6), we use three datasets, SCITE, FinCausal, and Organizational, that are more closely related to our data. In the future, this work can be expanded to pre-train the model on all these datasets, which may improve the performance.

This chapter introduced basic natural language processing concepts and concentrates mainly on information extraction. It also summarized the work that is related to information extraction. Then it discussed the literature behind the causality extraction from text and the datasets created as part of a challenge or for an individual study that is publicly available for users. Most of the datasets listed are publicly available except the Inter-Sentence Relation Extraction Dataset [91].

Table 2.1: Summary of a few of the available datasets on causality extraction and relation extraction, which includes some commonly used data and datasets that represent document-level relations, inter-sentence level relations, and implicit and explicit relations. This is not a complete list of available datasets, but this table gives an idea that several datasets for information extraction with different causal units are available. Different levels of information extraction from text (sentence level, document level, within sentence level, across sentence level) are available. From this, we use three of the datasets, SCITE, FinCausal, and Organizational which are more similar to our data for the transfer learning task in Chapter 6.

Dataset name	**Article link**	**Data link**	**Dataset size**
SemEval 2010 Task8 [86]	https://aclanthology.org/S10-1006.pdf	https://www.kaggle.com/datasets/drtoshi/semeval2010-task-8-dataset	10717 instances, 8000 train, 2717 test
PDTB-2.0 [87]	http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf	https://catalog.ldc.upenn.edu/LDC2008T05	40600 discourse relations
The TAC Relation Extraction Dataset (TACRED)[88]	https://aclanthology.org/D17-1004.pdf	https://catalog.ldc.upenn.edu/LDC2018T24	119,474 examples
Document-Level Relation Extraction Dataset (DocRED) [89]	https://aclanthology.org/P19-1074.pdf	https://github.com/thunlp/DocRED	5053 documents Human annotated
Automatic Content Extraction (ACE) [94]	https://www.ldc.upenn.edu/collaborations/past-projects/ace/papers-and-presentations	https://catalog.ldc.upenn.edu/LDC2006T06	Chinese, Arabic and English
Inter-Sentence Relation Extraction Dataset [91]	https://aclanthology.org/L18-1246.pdf		31,970 sentence pairs involving 17 different relations
Causal News Corpus [93]	https://aclanthology.org/2022.lrec-1.246.pdf	https://github.com/tanfiona/CausalNewsCorpus/tree/master/data/V2	2754 causal relations.
Organizational data [11]	https://www.mdpi.com/2078-2489/14/7/367	https://github.com/GoPeaks-AI/text2causalgraph	2234 sentences.
SCITE [70]	https://www.sciencedirect.com/science/article/abs/pii/S0925231220316027	https://github.com/Das-Boot/scite&5236sentences	5236 sentences.
FinCausal [95]	https://aclanthology.org/2022.fnp-1.16/	https://wp.lancs.ac.uk/cfie/fincausal2020/	4386 training and 265 validation data.

CHAPTER 3: CLINICAL PRACTICE GUIDELINES

A Clinical Practice Guideline (CPG) document serves as a roadmap for physicians, aiding in their decision-making process about diagnosis, management, and treatment. Clinical practice guideline is defined by the Institute of Medicine (IoM) as "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" [97].

Based on the above definition, guidelines have two parts:

- The basis for a Clinical Practice Guideline is a systematic evaluation of the research evidence related to a specific medical query. The emphasis is on the robustness of the evidence that informs the clinical decision-making process for that particular condition.

- A suite of suggestions that include both the evidence and assessments of the merits and disadvantages of various care alternatives. These recommendations address how patients with a particular condition should be treated, assuming all other factors are equal.

In 2011, the Institute of Medicine updated the definition of Clinical Practice Guidelines [98] as "Clinical Practice Guidelines are statements that include recommendations intended to optimize patient care that is informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options." They also define a set of criteria that can be used to decide the trustworthiness of the guidelines. They are listed below.

- The guidelines should be developed following a systematic evaluation of available evidence.

- The guidelines should be created by a team of proficient individuals, encompassing a diverse panel of experts. It's crucial that the panel also includes representations from the key groups affected.

- It's important to take into account significant patient subgroups and patient preferences appropriately.

- The establishment of the guidelines should be carried out in a clear and straightforward manner, focusing on minimizing any misleading representations, prejudices, or conflicting interests.

- There should be a clear portrayal of the various care alternatives and health outcomes. Additionally, the evidence quality and the strength of the suggestions should be assigned a rating.

- The guidelines should be reevaluated and updated as new evidence emerges that necessitates changes.

There are over 37,000 guidelines indexed in PubMed [1] as "clinical practice guidelines." In the past five years, approximately 50,000 documents mentioning gestational diabetes. Automated analysis is necessary (for various applications like comparing differences in guidelines, diagnostic support, etc.)

Given below is an example from the clinical practice guidelines, which are causal.

Example 3.0.1 *Inhaled steroids are the most effective preventer drug for adults and children for achieving overall treatment goals. Inhaled steroids are the recommended preventer drug for adults and children for achieving overall treatment goals* [99].

Example 3.0.1 from asthma guidelines has an obvious causality with inhaled steroids as the cause, and the effect is achieving treatment goals. (Implicit effect: Preventing asthma).

[1] `https://pubmed.ncbi.nlm.nih.gov/`

Example 3.0.2 *The decision to start screening mammography in women prior to age 50 years should be an individual one. Women who place a higher value on the potential benefit than the potential harms may choose to begin biennial screening between the ages of 40 and 49 years* [100].

In example 3.0.2 from breast cancer screening guidelines [100], the causality is not obvious. Here the modal verb "may" decides the causality. If there is a potential benefit, then biennial screening is recommended, which is a causal statement. This is an example of a vague causality because of the presence of a modal verb.

3.1 Clinical Decision Support System (CDSS)

Clinical practice guidelines (CPGs) are the primary source of knowledge for the CDSS. It includes a variety of tools and interventions. It can be computerized (ClinicalKey) or non-computerized (clinical guidelines) [101]. They are the tools for information management. The CDSS can be categorized based on various aspects such as "system function, a model for providing recommendations, communication style, underlying decision-making process, and human-computer interaction" [101]. Based on the characteristic "system function" [101] distinguishes CDSS into two functions

- Determining what is true? Which include diagnostic CDSSs like WebMD etc. (diagnosis websites).
- What to do? For the purpose of differential diagnosis, which tests should be ordered?

3.1.1 Work related to automatic information extraction from Clinical Practice Guidelines (CPG)

This section summarizes the prior work related to information extraction on the Clinical Practice Guidelines. However, there is no existing work on causality extraction from CPGs.

There are plenty of works that discuss information extraction from Clinical Practice Guidelines. Extracting clinical findings from the outpatient progress notes was done by [102]. This work uses the CPGs to compare it against the chart notes. A chart note is deemed relevant if it can be assessed against the National Heart, Lung, and Blood/National Asthma Education program's clinical practice guidelines. This project aims to select outpatient notes that can be evaluated, ascertain the patient's requirement for inhaled anti-inflammatory medications, and measure the severity of asthma. This initial project contrasts their findings with the judgment of an expert panel of practitioners. By utilizing a PERL-based partitioning program, they identified the need for anti-inflammatory medications 0.76 of the time.

A few years ago, [103] targeted the automated extraction of diagnosis and treatment procedures from these guidelines. Experts annotated the guidelines for diagnosing and treating chronic heart failure (CHF), as released by the European Society of Cardiology. After annotation, 171 sentences were identified. The referenced work recognized the entities and extracted the predictions using SemRep. OpenNLP and Stanford parsers were then utilized for shallow syntactic parsing, followed by Named Entity Recognition (NER). From the NER, patient context extraction was carried out using Regular Expressions (RE) with an F1-score of 0.75. In text pre-processing, they achieved an F1-score of 0.91, and SemRep correctly mapped 87% of the entities. Another similar work [104] introduces a method for automatically collecting useful information using rules rooted in both syntactic and semantic information. These information extraction rules are defined by the semantic relations derived from the UMLS. The study finds that rules formed using semantic and syntactic information prove beneficial and valuable. For evaluation, they use the "management of labor" clinical practice guideline, achieving recall and precision values of 0.69 and 0.79, respectively. A pattern-based approach was used by [105], which contrasts a manually developed ontology for CPG eligibility criteria with a top-level ontology stemming

from a semantic pattern-based approach. They design a model that amalgamates knowledge representation and knowledge acquisition technologies. The model they have created incorporates both the conceptual and procedural knowledge required for parsing. It also includes modules for coding a CPG.

A recent work [106] introduces an innovative system that blends the methodologies of Natural Language Processing (NLP) and Fuzzy Logic. Initially, a semantic extraction from the medical guidelines was performed using Long Short-Term Memory (LSTM). Subsequently, fuzzy rules were developed using the extracted semantics. These fuzzy rules are capable of identifying new cases in the learning domain. It also predicts and extracts the grade of recommendation. Its performance was compared with the latest NLP techniques for clinical information extraction. In the initial phase of the approach, the F1-score for the NLP-Fuzzy system is 0.88. For the prediction of the grade of recommendation using the fuzzy rules, the F1-score is 0.97. The machine learning methodology was used by another similar work [107] to extract and categorize Conflicts Of Interest (COIs) from disclosure statements indexed in PubMed. The correlation between the aggregate COI in medical literature and drug safety as determined by the FDA is assessed in the study. For each drug, they gathered the 10,000 most pertinent articles and cross-referenced them with the COI database. Spacy NER, trained on English web text, is employed for a small sample of COI disclosure statements (100 samples). The statements were then manually corrected, leading to a 0.25 enhancement in named entity recognition accuracy. Given that the data is scattered, a quasi-Poisson regression was employed for the analysis. Incidence Rate Ratios (IRR) with 0.95 confidence intervals were calculated for each association between COIs and adverse events.

Recently, Large Language Models (LLMs) have been employed for a variety of NLP tasks, including those involving information extraction. LLMs can be fine-tuned to cater to a specific dataset, or a prompt-based approach can be utilized. An illustrative

study [108] measures the efficiency of the few-shot learning performance of GPT-3 in tasks related to text classification and information extraction. A recent survey article [109] provides a summary of the methods and solutions employed for information extraction. It also highlights the challenges encountered when extracting information from medical documents.

3.2 Existing works and datasets on causality extraction from medical text

The previous section introduced the Clinical Practice Guidelines (CPGs) and summarized some of the prior art that is related to information extraction from CPGs. This chapter discusses some of the available work and datasets that are related to causality extraction from medical text.

3.2.1 Causality extraction from the medical text – Related work

This section discusses the work related to causality extraction from medical text. Among the works that are summarized below, [110], [111], and [112] are more closely related to the task discussed in this work. However, all this work only concentrates on identifying condition-action statements from the CPGs. They are not extracting cause-effect relations or the signals from the CPGs, which we are doing. Also, the domain adaptability across domains and extracting the degree of influence are not done in the above-mentioned related works. [113] extracts recommendation statements from CPGs, which is less related to this work even though it does an information extraction from CPG. [114], [115] are less related to our work because they are doing different tasks on CPGs.

A few years ago, [111] aimed to automatically detect conditional activities in medical guidelines using a rule-based approach. They calculate a score based on the presence of trigger words such as "if" and "should" and sequences of semantic types from the UMLS. Following this approach with the asthma guidelines, they achieved precision, recall, and F1-score values of 0.88, 0.75, and 0.81, respectively. Similar

works on detecting condition action statements from CPGs, [110], [112] apply supervised machine learning techniques to classify sentences according to whether they express conditions and actions. The study utilizes a manually annotated medical guideline on Asthma, Rhinosinusitis, and Hypertension as the gold standard data. Candidate condition-action sentences are extracted using regular expressions applied to POS tags. Classifiers such as ZeroR, Naive Bayes, J48, and Random Forest, were deployed for the research. For the asthma guidelines, the Naïve Bayes Classifier recorded superior performance with an F1-score of 0.33. The Random Forest Classifier delivered an F1-score of 0.535 for the Rhinosinusitis guidelines and an F1-score of 0.587 for the Hypertension guidelines. Another study by Hussain et al. [113] uses heuristic patterns to identify recommendation statements in Clinical Practice Guidelines (CPG). In this process, meaningful data known as recommendation statements are pulled from the CPGs, restructured into a word vector format and used to train an ensemble learning model. This model involves several methods, including Naive Bayes, Generalized Linear Model, Random Forest, Deep Learning, and Decision Tree, and its primary function is to discern whether a sentence is a recommendation statement. In terms of results, the Naive Bayes model achieved an F1-score of 0.82, the Generalized Linear Model scored 0.74, the Deep Learning model reached 0.66, the Decision Tree model achieved 0.66, the Random Forest classifier scored 0.74, and the Ensemble Learner attained a score of 0.83.

A review article [116] documents the existing methods and tools for clinical concept extraction. A literature search was conducted to gather electronic health records and published articles. The retrieved articles underwent screening and selection. Documents retrieved automatically were also manually screened. Of the 10,441 total articles retrieved, there were 6,686 unique ones. These articles were manually filtered, leaving 228 articles that mainly focused on clinical concept extraction. These chosen articles were then overviewed in the article.

A study by Schlegel et al. [115] introduces a framework known as Clinical Tractor, which employs Natural Language Understanding (NLU) techniques to generate semantic representations of entire Clinical Practice Guidelines (CPG). The process begins by converting each guideline document into an XML format, followed by text processing on the XML data. This processed text then undergoes dependency parsing and Named Entity Recognition. A tool called Propositionalizer is used to consolidate annotations with identical start and end positions. The Background Knowledge Alignment System (BKAS) is then employed to compare the text with lexical resources such as WordNet and VerbNet and to identify Ontological terms. These terms are then imported into a knowledge base (KB). Finally, a syntax-to-semantics mapper is applied, resulting in the creation of a semantic KB.

Xie et al., [114] uses pre-trained language models to produce an extractive summarization of biomedical literature. The Sci model is trained via the EBM-NLP dataset for a sequence labeling task, the goal of which is to pinpoint PICO (Population, Intervention, Comparison, and Outcome) components within the text. These PICO elements, brimming with domain knowledge, are then incorporated into the pre-trained language models for the extractive summarization, achieved by training a blend of generative and discriminative adapters. The effectiveness of this method is tested on the CORD-19 dataset and the PubMed-Long dataset. In terms of results, a rogue1 F1-score of 32.04 was achieved on the CORD-19 dataset - PubMed-all-full with all adapters, while the PubMed-Long dataset yielded a Rogue1 F1-score of 36.39. A more recent study [117] explores the use of ChatGPT for clinical text mining, specifically for extracting structured data from unstructured healthcare texts and focusing on biological named entity recognition and relation extraction. Initial results showed that using ChatGPT directly for these tasks was not effective. To address these issues, we suggest a new training approach that involves creating a large amount of high-quality synthetic data with labels using ChatGPT and fine-tuning a local model for

the task at hand. It also reduces the time and effort for data collection and labeling.

3.2.2 Causality extraction from medical text - Datasets

This section summarizes the causality extraction datasets created with the medical text. These datasets can be a part of a challenge, or they can be developed for individual research. A summary of details about the datasets is available in Table 3.1.

The National NLP Clinical Challenges (n2c2) - Cohort Selection and Adverse Drug Events & Medication Extraction [118] is a competition centered on extracting information about adverse drug events from clinical records. Participants are given the raw text of discharge summaries and can choose any model for the task of relation extraction. The data set comprises 505 discharge summaries sourced from the MIMIC-III (Medical Information Mart for Intensive Care-III) clinical care database [119]. Each record is searched for the keyword 'ADE' in the international classification of diseases code description and manually checked to confirm the presence of an Adverse Drug Event. Each document is then annotated by two annotators, with a third resolving any disagreements. Out of the 505 discharge summaries, 303 were used for training and the remainder for testing.

The 2016 Centers of Excellence in Genomic Science (CEGS) Neuropsychiatric Genome-Scale and RDOC Individualized Domains (N-GRID) challenge [120] is centered on predicting symptom severity from neuropsychiatric clinical records. The data for this challenge is provided by Partners Healthcare and the N-GRID project of Harvard Medical School, comprising a total of 816 neuropsychiatric clinical records. Of these, 600 records are used for training, and the remaining 216 for testing. Out of the 600 training records, 325 were annotated by two annotators, 108 were annotated by a single experienced annotator, and 167 were left unannotated. The inclusion of unannotated records in the training data is intended to provide participants with experience in a semi-supervised approach.

The Medication, Indication, and Adverse Drug Events (MADE) 1.0 corpus [121]

was developed for the MADE 2018 challenge. The goal of this challenge is to extract information about medication, indications, and adverse drug events from Electronic Health Record (EHR) notes. The corpus comprises 1,089 fully de-identified longitudinal EHR notes from 21 randomly selected cancer patients at the University of Massachusetts Memorial Hospital. Three shared Natural Language Processing (NLP) tasks were designed based on this corpus. The first task, Named Entity Recognition (NER), aims to identify medications and their attributes (dosage, route, duration, frequency), indications, adverse drug events, and severity. The second task, Relation Identification (RI), seeks to identify the relationships between named entities, specifically medication-indication, medication-ADE, and attribute relations. The third task evaluates NLP models that perform both NER and RI tasks jointly. The annotation guidelines were formed through an iterative process, adapting and extending from the 2009 i2b2 shared task of the Medication Challenge annotation guideline. It defines nine named entity types and seven relation types. Two annotators performed the annotation, and Fleiss' Kappa measure of inter-annotator agreement was used on three documents from the corpus annotated by five annotators.

The Cantemist (CANcer TExt Mining Shared Task) study [122] involves the development of resources for named entity recognition, concept recognition, concept normalization, and clinical coding, all with a focus on cancer data in Spanish. Named entity recognition involves automatically identifying tumor morphology mentions in text. Concept recognition and normalization require the model to identify all tumor morphology entity mentions and their corresponding eCIE-O codes. Clinical coding requires the model to provide a ranked list of correct eCIE-O code assignments for each document. The Cantemist corpus comprises 1301 oncological clinical case reports in Spanish, with 16030 annotations, 850 unique codes, and 63016 sentences. The training subset includes 501 documents, the development phase includes 500 documents, and the test subset includes 300 documents. All documents were manually

annotated by clinical experts for mentions of tumor morphology. The Cantemist test set was released along with an additional collection of 4,932 clinical case documents.

The BioCreative V CDR (BC5CDR) corpus [123] was developed by a team of Medical Subject Headings (MeSH) indexers for disease/chemical entity annotation, while the Comparative Toxicogenomics Database (CTD) curators performed the CID relation annotation. To ensure high-quality annotation, detailed guidelines, and automatic annotation tools were provided. The BC5CDR corpus comprises 1500 PubMed articles, with 4409 annotated chemicals, 5818 diseases, and 3116 chemical-disease interactions. Each entity annotation includes the mention of text span and a normalized concept identifier annotated using MeSH vocabulary. To ensure annotation accuracy, two annotators independently annotate the entities, followed by a consensus annotation. The training, development, and most of the test set were randomly selected from the CTD Pfizer corpus, which includes over 150,000 chemical-disease relations from 88,000 articles.

BioCause [124] is a dataset specifically designed for the extraction of causality in the biomedical field. It involves manual annotation of causality on existing bio-event corpora. The dataset includes 19 open-access biomedical journals from the infectious disease subdomain, annotated to include 851 causal relations.

All the datasets listed in Table 3.1 are available publicly except [125], which is related to our work. The annotation guidelines discussed in this thesis are partly adopted from BioDRB [125]. However, the BioDRB is not annotated on the CPGs. Another dataset [110] is related to our work, but it concentrates only on the condition-action statements from the CPGs. [118], [121] is an adverse drug event extraction task, [120] extracts symptom severity. They are less related to our work even though they extract information from the medical text because they aim to extract different types of information, which is not a concern of this thesis.

From the summary of the medical relation extraction datasets, we can understand

Table 3.1: Summary of few of the available datasets on causality extraction and relation extraction of medical text. From this summary of datasets, we can understand that a lot of datasets are available for relation extraction from medical text. Most of them extract different information like adverse drug event etc but only a very few of them are causality extraction dataset and no dataset is annotated on CPGs.

Dataset name	**Article link**	**Data link**	**Dataset size**
National NLP Clinical Challenges (n2c2)- Cohort Selection and Adverse Drug Events & Medication Extraction [118]	https://academic.oup.com/jamia/article-abstract/27/1/3/5581277?redirectedFrom=fulltext&login=false	https://n2c2.dbmi.hms.harvard.edu/data-sets	505 discharge summaries drawn from the MIMIC-III database
2016 Centers of Excellence in Genomic Science (CEGS) Neuropsychiatric Genome-Scale and RDOC Individualized Domains (N-GRID) challenge [120]	https://www.sciencedirect.com/science/article/pii/S1532046417300874	https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/	816 de-identified neuropsychiatric clinical records
Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0) [121]	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6860017/pdf/nihms-1518894.pdf	https://bio-nlp.org/index.php/announcements	1092 medical notes from 21 randomly selected cancer patients'
CANTEMIST CANcer TExt Mining Shared Task – tumor named entity recognition [122]	http://ceur-ws.org/Vol-2664/cantemist_overview.pdf	https://temu.bsc.es/cantemist/?p=4338	1900 clinical cases
BC5CDR -BioCreative V CDR task corpus [123]	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4860626/pdf/baw068.pdf	https://github.com/JHnlp/BioCreative-V-CDR-Corpus	1500 PubMed articles with 4409 annotated chemicals, 5818 diseases and 3116 chemical-disease interactions.
The biomedical discourse relation bank- BioDRB [125]	https://link.springer.com/content/pdf/10.1186/1471-2105-12-188.pdf		24-articles from GENIA corpus. 112,000-word tokens and approximately 5000 sentences.
Identifying Condition-Action Statements in Medical Guidelines Using Domain-Independent Features [110]	https://arxiv.org/pdf/1706.04206.pdf	https://data.world/hematialam/condition-action-data	Hypertension 2014, Chapter 4 of 2008 asthma, 2012 rhinosinusitis guideline
BioCause: Annotating and analysing causality in the biomedical domain [124]	https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-2	http://www.nactem.ac.uk/biocause/	19 articles of infectious diseases - 851 causal relations.
MediCause: Causal Relation Modelling and Extraction from Medical Publications [34]	https://ceur-ws.org/Vol-3184/TEXT2KG_Paper_1.pdf	https://github.com/IReklos/MediCause	1202 annotated examples.

that there are a lot of datasets available. But only a very few are causality extraction datasets [34], [124], [126], [127]. Notably, none of these datasets are annotated on CPGs. This highlights a gap in the field and underscores the need for either a transfer learning approach or a specifically annotated corpus. This thesis addresses both these needs.

This chapter provided an introduction to Clinical Practice Guidelines. It then summarized some of the previous work on information extraction from these guidelines. Additionally, it reviewed prior research on causality extraction and the datasets currently available for this purpose. The chapter concludes by identifying the existing gap in the field of causality extraction from Clinical Practice Guidelines and outlines how this thesis aims to address these gaps.

CHAPTER 4: CAUSALITY: DEFINING THE MAIN CONCEPTS AND DATA ANNOTATION

In this chapter, the two primary research areas of this thesis - causality and causality extraction - will be explored. We informally introduced causality in Chapter 1, and in this Chapter, we introduce causality formally with examples. Each of the terms is introduced with an example from the medical guidelines. Additionally, the causal unit that shall be utilized during the annotation process will be precisely defined.

4.1 Causality extraction

Causality represents the relationship between the "cause" and "effect" in text. The occurrence of a cause will trigger an effect to happen. The process of automatically extracting such relations is *causality extraction.*

Causality can be within a sentence, or it can be across a sentence. If it is within a sentence, it is called *intra sentence relation*, and if the relationship is across the sentences, then it is called *inter sentence relation.*

Example 4.1.1 ***Intra sentence relation:***

Women aged 45 to 49 years should be screened with mammography annually [100].

Condition: *age 45 to 49,*

Action: *screened with mammography annually,*

Effect: *detect breast cancer.*

In the above example from breast cancer screening guidelines [100], we have a condition action statement which is a direct condition/ action statement.

Example 4.1.2 ***Inter sentence relation:***

In children under 5 years, higher doses may be required if there are problems in ob-

taining consistent drug delivery. Titrate the dose of inhaled steroids to the lowest dose at which effective control of asthma is maintained [99].

Condition: *children under 5 years,*

Cause: *if there are problems in obtaining consistent drug delivery,*

Action: *higher doses may be required. Titrate the dose of inhaled steroids to the lowest dose,*

Effect: *control asthma.*

The above example from the asthma guidelines [99] is a condition, cause, and action statement. Here, the action is not direct but depends on the modal verb "may." If there are problems with consistent drug delivery, higher doses may or may not be required.

4.2 Causal unit

Different causal units have been defined by different works [128], [33], [32, 129]. We define our causal unit based on [125] and [130]. Below is the definition of our causal unit for organizational and medical data; this is not the only way to define it.

4.2.1 Causal unit in Organizational data

- Cause: Give two arguments, `arg1` and `arg2`, if they are related causally and are not in a conditional relation.
- Effect: Possible effect that the condition/cause/action will lead to.
- Causal Trigger (CT): Represents the link between cause and effect.

4.2.2 Causal unit in medical data

- Cause: Give two arguments, `arg1` and `arg2`, if they are related causally and are not in a conditional relation.
- Condition: The condition which is true and will lead to a solution. It represents conditional relations.

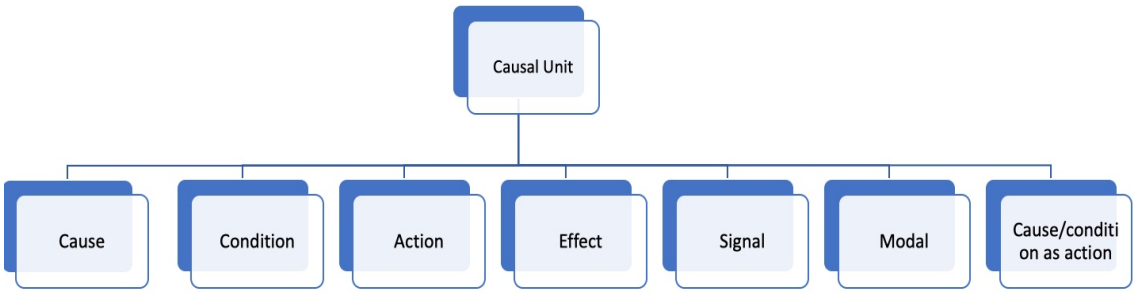


Figure 4.1: Tree showing the causal unit. If there is a modal verb in that sentence, it will be annotated. In some cases, cause/condition can also act as an action; in those cases, both the senses will be annotated.

- Action: Possible action that needs to be taken for a condition/cause. This depends on the condition/cause.
- Effect: Possible effect that the condition/cause/action will lead to.
- Signal: Represents the link between the cause and effect.

There can be sentences with condition-action, condition-effect, cause-effect, cause-action, etc., or all of them. In Organizational data, the link between cause and effect is given by a causal trigger, whereas in medical text, we define the link between cause and effect as a signal as given in [130]. Throughout this thesis, we will use both these terms interchangeably.

Table 4.1: Summary of the causal unit defined in this work with examples from medical guidelines. In the examples, a Cause is marked as C, a Condition is marked as CO, an Effect as E, an Action as A, a Signal as S, and a Modal verb as MD. The cause, condition, effect, action, or signal identified in the sentence are marked inside brackets. In some cases causes can be effects and vice versa.

Causal unit	**Example**
Cause	Some [pregnant persons]$_S$ are [screened earlier than 24 weeks of gestation]$_A$ because they [have risk factors for type 2 diabetes, such as obesity, family history of type 2 diabetes, or fetal macrosomia during a previous pregnancy]$_{C,E}$
	[Potential harms]$_S$ of [screening for gestational diabetes]$_C$ include [psychological harms (anxiety, depression), intensive medical interventions (induction of labor, cesarean delivery, or admission to the neonatal intensive care unit [NICU]), and negative hospital experiences related to labeling (reduction in breastfeeding and fewer newborns staying in the mother's room)]$_E$ that may be associated with a diagnosis of gestational diabetes. Possible adverse effects of treatment include neonatal or maternal hypoglycemia, increased risk of small for gestational-age infants, and maternal stress.
	AAFP acknowledges that diabetes, like all chronic conditions, exists on a continuum and that insulin resistance is a [risk factor]$_S$ for developing diabetes, the [current evidence does not show improvement in long-term health outcomes]$_C$ for [screening for prediabetes in adults who have obesity or overweight]$_E$.
	Many [studies]$_S$ [comparing different inhaled steroids]$_C$ are of [inadequate design and have been omitted from further assessment]$_E$.
Condition	[You'd get fatter]$_E$ if [you ate too much]$_C$
	There is [limited evidence]$_S$ that [screening with mammography reduces breast cancer mortality]$_E$ in [women 40- 49 years of age]$_C$.
	If he had [reduced his sugar intake]$_C$, he would be [free from diabetes]$_E$.
Modal	[Pregnant persons with gestational diabetes]$_C$ are at [increased risk for maternal and fetal complications]$_E$ and [may]$_{MD}$ benefit from [early identification and treatment]$_A$.
Cause/ Condition is an action	In the [SHOCK trial]$_C$, [patients were randomized to medical therapy or emergency revascularization]$_A$. Among the [patients randomized to revascularization]$_C$, [two-thirds of patients were referred for PCI and one-third for CABG]$_A$, and the decision to proceed with PCI or CABG was made by the treating physician.

The causal units defined in this thesis, along with examples from medical guidelines, are given in this section.

1. ***Cause:*** is defined between arguments if they are related causally. Examples of causes are given below.

   Example 4.2.1 *Some* $[$*pregnant persons*$]_S$ *are* $[$*screened earlier than 24 weeks of gestation*$]_A$ *because they* $[$*have risk factors for type 2 diabetes, such as obesity, family history of type 2 diabetes, or fetal macrosomia during a previous pregnancy*$]_{C,E}$ [131].

   Example 4.2.2 $[$*Potential harms*$]_S$ *of* $[$*screening for gestational diabetes*$]_C$ *include* $[$*psychological harms (anxiety, depression), intensive medical interventions (induction of labor, cesarean delivery, or admission to the neonatal intensive care unit [NICU]), and negative hospital experiences related to labeling (reduction in breastfeeding and fewer newborns staying in the mother's room)*$]_E$ *that may be associated with a diagnosis of gestational diabetes. Possible adverse effects of treatment include neonatal or maternal hypoglycemia, increased risk of small for gestational-age infants, and maternal stress* [131].

   Example 4.2.3 *AAFP acknowledges that diabetes, like all chronic conditions, exists on a continuum and that insulin resistance is a* $[$*risk factor*$]_S$ *for developing diabetes, the* $[$*current evidence does not show improvement in long-term health outcomes*$]_C$ *for* $[$*screening for prediabetes in adults who have obesity or overweight*$]_E$ [132].

   Example 4.2.4 *Many* $[$*studies*$]_S$ $[$*comparing different inhaled steroids*$]_C$ *are of* $[$*inadequate design and have been omitted from further assessment*$]_E$ [99].

2. ***Condition:*** which describes all the conditional relations. Conditions were annotated as CO.

   Example 4.2.5 $[$*You'd get fatter*$]_E$ *if* $[$*you ate too much*$]_{CO}$

   Example 4.2.6 *There is* $[$*limited evidence*$]_S$ *that* $[$*screening with mammography reduces breast cancer mortality*$]_E$ *in* $[$*women 40- 49 years of age*$]_{CO}$ [100].

   Example 4.2.7 *If he had* $[$*reduced his sugar intake*$]_{CO}$*, he would be* $[$*free from diabetes*$]_E$*.*

3. ***Modal:*** If there is a modal verb like may, must, can, could, etc., it was annotated as MD.

   Example 4.2.8 $[$*Pregnant persons with gestational diabetes*$]_C$ *are at* $[$*increased risk for maternal and fetal complications*$]_E$ *and* $[$*may*$]_{MD}$ *benefit from* $[$*early identification and treatment*$]_A$ [131].

   Here the action for a particular cause may or may not help to reduce the effect.

4. ***Cause/Condition is an action:*** The situations where the cause/condition will also be an action or vice versa.

   Example 4.2.9 *In the* $[$ *SHOCK trial* $]_C$*,* $[$ *patients were randomized to medical therapy or emergency revascularization* $]_A$*. Among the* $[$*patients randomized to revascularization*$]_C$*,* $[$*two-thirds of patients were referred for PCI and one-third for CABG* $]_A$*, and the decision to proceed with PCI or CABG was made by the treating physician* [133].

In the above examples, the phrases marked with C indicate a Cause, phrases with CO indicate a condition, phrases marked with E represent an Effect, A represents an Action, and S indicates a Signal. A summary of all the causal units and their examples are available in Table 4.1

4.3 Degree of influence of cause to effect

The degree of influence represents a cause's influence on the effect.

Example 4.3.1 *Risk analysis shows that all women diagnosed at or before age 50 and treated with breast-conserving therapy have a 20% or higher lifetime risk for a new breast cancer* [134].

The above example is from the breast cancer screening guideline by the American College of Radiology(ACR) [134]. Here the degree of influence from cause (women diagnosed at or before age 50 and treated with breast-conserving therapy) to effect (higher lifetime risk for new breast cancer) is 20% or higher.

4.4 Data Annotation

This thesis incorporates two sets of annotations - organizational data and medical data. This section provides an overview of the data and the annotation process, both of which have undergone manual annotation. For the organizational data, we manually evaluated the concordance among the annotators. Meanwhile, for the medical data, we computed the agreement through an automated process.

4.4.1 Organizational data

As we explained in [11],

> The organizational data was manually annotated on the 2020 SEC 10-K Documents of 65 S&P Financial Companies. Five graduate students trained in business analytics, business administration, and/or economics were hired to manually annotate the causal insights from the documents based on a predefined dictionary of causal trigger words (or cue phrases). Causal triggers are also called causal trigger words. At least two students carefully read each sentence with trigger words to ensure it described a cause-and-effect relationship. As is customary in the area of causality

extraction [20, 135], only sentences with trigger words were considered causal. Thus, the article addresses explicit causality. However, this is not a major limitation since causal sentences without causal triggers are relatively rare. Two thousand two hundred thirty-four causal sentences were identified and manually annotated. For each of the identified causal sentences, the cause/effect relationship was marked using tags. In the manual annotation, the effects were marked with the <outcome> tag. Five graduate students manually annotated causes, triggers, and outcomes. After one round of discussions to resolve disagreements, there was a 100% inter-rater agreement. However, the 100% agreement did not imply complete consistency, e.g., some phrases included the determiner 'the' in some sentences but omitted it in others. An example sentence is given below:

Example 4.4.1 *<causal-relation> When a <cause> policyholder or insured person becomes sick or hurt </cause>, the Company <trigger> pays </trigger> <outcome> cash benefits fairly and promptly for eligible claims </outcome> </causal-relation>.*

4.4.2 Medical data

We annotated seven documents of gestational diabetes guidelines from various societies like the American Diabetes Association (ADA) [136], US Preventive Services Task Force (USPSTF) [131], American College of Obstetrics & Gynecology (ACOG) [137], American Academy of Family Physician (AAFP) [132], Endocrine Society [138] etc. This idea to annotate the gestational diabetes guidelines is based on an opinion from Dr. Luciana Garbayo, who is an assistant professor of Philosophy and Medical Education at the University of Central Florida College of Medicine. She suggested the presence of causal sentences in the gestational diabetes guidelines.

Two annotators were recruited. Given a text document, their task is to read the

document and mark the cause, effect, condition, action, modal, and degree of influence with tags. The cause will be marked as C, effect as E, signal as S, condition as CO, action as A, modal as MD, and degree of influence as DI. The phrases containing any of these causal phrases should be differentiated. For example, the beginning of a cause phrase will be marked as <C> and the end as </C>.

4.4.3 Inter-annotator agreement for the medical data

Due to the intricacy of causality extraction, which involves annotators labeling varying text spans as "cause," "effect," and so on, computing agreement between two annotators can be challenging as it requires comparing two spans of texts. Traditional methods of inter-annotator agreement, such as the Kappa statistic [139], are inadequate due to their need for classifications to fit into mutually exclusive and discrete categories. Therefore, we decided to assess agreement using both exact match and relaxed match criteria. The F-measure is used for the exact match [140, 141, 124]. In the case of the relaxed match, the average distance between phrases is computed. Initially, the annotated phrases, their corresponding labels, and the full sentence they are derived from are extracted from the entire annotated document. These annotations, originating from both annotators, are then compared and amalgamated based on the sentence. The resulting merged table thus features the sentence, the extracted phrase, and the labels as marked by Annotator 1 and Annotator 2. In total, 514 matching phrases have been identified. An overall agreement computed as a Jaccard similarity of 0.66 was obtained. Details of the inter-annotator agreement computation are given below.

From the merged data table, the inter-annotator agreement was computed. This is done by computing the match between the annotations as follows.

- Relaxed match – Both annotator's phrases overlap with each other but are not necessarily an exact match.

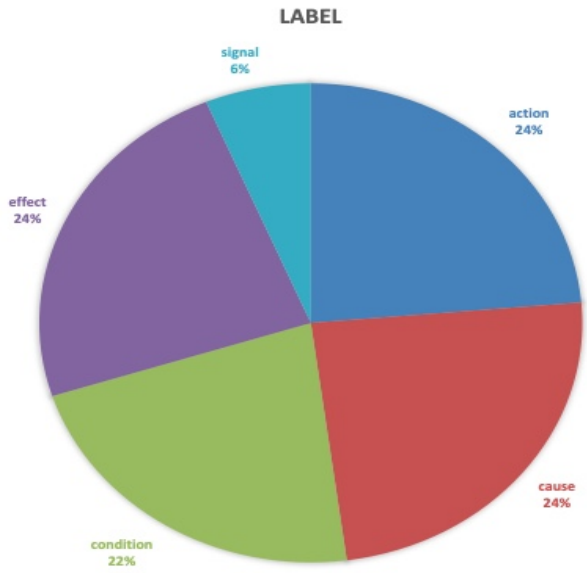


Figure 4.2: Distribution of the labels in the corpus. The percentage of almost all the labels is around 24% except signal, which is only 6%

- Exact match – Both annotator's phrases exactly match.

To execute the relaxed match, we employ the Levenshtein distance [142] and the Jaccard distance [143]. The Levenshtein distance quantifies the difference between two string sequences, indicating the minimum single-character edits required to transform one word into another. Jaccard similarity computes the degree of relatedness between two finite samples by dividing the intersection's size by the size of the sample sets' union. The Jaccard distance is subsequently calculated by subtracting the Jaccard similarity from 1. The Python library Levenshtein [1] is used in computing the Levenshtein distance. The Jaccard index was computed using the Python library textdistance[2]. The Levenshtein distance and the Jaccard distance between the annotators are summarized in Table 4.2

From Table 4.2, we can understand that there is an average Levenshtein distance of 0.41 and an average Jaccard distance of 0.34. In most cases, both annotators annotated the same sentence with the same labels, but the length of the phrase was

[1] `https://pypi.org/project/python-Levenshtein/`
[2] `https://pypi.org/project/textdistance/`

Table 4.2: Relaxed match between the annotated phrases. Levenshtein distance is the minimum number of edits required to transform one phrase to another, whereas Jaccard distance is the amount of non-overlap between phrases. The lower the distance, the agreement is higher. The distance is higher for action. In most of the cases where there is a mismatch, the length of the phrase by both the annotators was different.

	Levenshtein distance	Jaccard distance
Cause	0.22	0.27
Condition	0.34	0.21
Effect	0.37	0.31
Action	0.87	0.48
Signal	0.26	0.44

Table 4.3: For a given phrase, the labels annotated by annotators 1 and 2 are compared. An average F1 score of 0.69 was obtained. From the F1-score, we can understand that both the annotators agree on most of the categories except the signal for which the F1-score is low.

	Precision	Recall	F1-score
Cause	0.86	0.71	0.77
Condition	0.56	0.85	0.67
Effect	0.85	0.90	0.88
Action	0.89	0.70	0.78
Signal	0.25	0.62	0.36

different. The exact match between the phrases is computed by finding the exact string match between phrases 1 and 2. Out of the 514 phrases, 112 phrases are exact matches. The match between the labels for the same phrase by both annotators is also computed with an average F1 score of 0.69. The match between the labels for each subcategory is given in Table 4.3

In this chapter, the main concept of this thesis, causality, was introduced. Also, the process of annotation was discussed in detail.

CHAPTER 5: CAUSALITY EXTRACTION FROM ORGANIZATIONAL TEXT AND MEDICAL TEXT

This chapter will explain the experiments we did with two specific datasets, organizational data and medical data. The organizational data is annotated on a set of financial documents, whereas the medical data is annotated on a set of Clinical Practice Guidelines.

5.1 Causality extraction on the organizational data

This section will discuss the data preparation we did for the organizational data. It also explains the methods we use to classify the causal sentences and extract causalities from those sentences. The content and the results of this section are copied from our [11].

5.1.1 Data Preparation and Preprocessing

As we have explained already in our work on causality extraction [11],

> Initially, 65 reports for the year 2019 from the 10-K annual report documents of Standard & Poor (S&P 500) Financial Companies were retrieved from the Securities and Exchange Commission (SEC) website. On the SEC website, the data is in Inline (eXtensible Business Reporting Language) XBRL format (IXBRL) format. This text is extracted in JSON format without filtering using Trafilatura [144], a Python package, and cleaned the data using the NLTK package [145]. From this extracted and cleaned text, the causal sentences were identified using the causal trigger words. About two dozen trigger words and their variants (man-

ually) adopted from the [146] and from online sources[1] (last accessed on 15 June 2023) were used. If a trigger word was present in a sentence, it was marked as a possible causal sentence using tags <causal-relation>. A JSON dataset, with possible causal sentences marked, was given to a set of graduate students who read the sentences and marked which part of a sentence was cause/effect. They were marked using the tags given in Example 4.4.1.

The next step was to convert the annotated tags into the BIO label format. Whenever there was a cause tag, the beginning of that tag would be marked as the "B-C" beginning of the cause, and the rest of the cause was marked as "I-C" inside the cause. Similarly, the beginning of the effect would be marked as "B-E", the rest of the effect as "I-E", and the beginning of a causal trigger would be marked as "B-CT", and the rest of the causal trigger as "I-CT". The rest of the words, which were not cause/effect/trigger, were marked with "O" to indicate outside (i.e., not a label of interest). From the annotated tags, BIO labels were marked using regular expressions. The data in the BIO label format was simplified into the IO label format, which improved the consistency of annotations.

5.1.2 Causal sentence classification from Organizational data

As given in our work on causality extraction [11],

BERT [12] for sequence classification was fine-tuned on our dataset. We ran two sets of experiments to classify the causal sentences. For both sets of experiments, the positive sample consisted of manually annotated causal sentences. The differences lay in the negative samples. The first set

[1] http://web.mit.edu/course/21/21.guide/tran-cwp.htm, https://languageonschools.com/free-english-lessons/conjunctions/transition-words-cause-and-effect/, https://continuingstudies.uvic.ca/elc/studyzone/570/pulp/hemp5

of negative classes consisted of all cases containing a causal trigger where the sentence did not contain a causal relation (selected from organizational data) and a random sample of sentences without causal relations and without causal triggers (selected from Twitter data). The second set of negative samples consisted of all cases of a causal trigger where the sentence did not contain a causal relation. Our data had an equal number of positive and negative samples. The obtained data was divided into train and test data. In the first case, we obtained a macro F1-score of 0.89, and, in the second case, a 0.88 macro F1-score. The detailed results of the causal sentence classification for the two above-mentioned scenarios are given in Appendix A, Table A.1, and Table A.2.

5.1.3 Causality extraction from Organizational data

As we explained in [11],

> BERT is a state-of-the-art performing model for many NLP tasks, including relation extraction [147, 148]. We used the SpanBERT and DistilBERT models, adapted for token classification for causality extraction. Based on an 80% training set and 20% test set from the manually-annotated gold data, the performance of the SpanBERT model had a macro average F1-score of 0.89, macro average precision of 0.87, and macro average recall of 0.91 and DistilBERT had an average F1-score of 0.86, macro average precision of 0.81, and macro average recall of 0.91. From Table 5.1, it can be observed that Span BERT's performance was better for the causal triggers. However, DistilBERT performed slightly better for cause and effect.
>
> We also tried using the BERT-large model for the same causality extraction task. BERT-large obtained a macro average F1-score of 0.83, macro

Table 5.1: Summary of SpanBERT(Sp) and DistilBERT's(Dis) performance on the Organizational data for the causality extraction task (CE-ORG). Each token in the text was assigned a Cause (C), Effect (E), and Causal Trigger (CT) label. The results given above were obtained by splitting the manually annotated gold data into train and test partitions, from which the training partition was used to fine-tune BERT.

	P(Sp)	**R(Sp)**	**F1(Sp)**	**P(Dis)**	**R(Dis)**	**F1(Dis)**
Cause	0.82	0.86	0.84	0.78	0.93	0.85
Causal trigger	0.93	0.97	0.95	0.77	0.86	0.81
Effect	0.86	0.90	0.88	0.88	0.94	0.91

precision of 0.78, and macro recall of 0.90, which was lower than the performances of DistilBERT and SpanBERT. We have also tried a few-shot prompting on this data.

Finally, we noted that when using the BIO-label format to include the beginning and the inside tags for cause, effect, and trigger, we obtained a macro average F1-score of 0.60, accuracy of 0.73, macro average precision of 0.73, and macro average recall of 0.60 using DistilBERT.

In this work on Organizational data, we built a pipeline that will take the text document as input and output a knowledge graph. An algorithm[2] that explains the steps of the pipeline is given in Algorithm 1

[2]Algorithm copied from [11]

Algorithm 1 Text to Knowledge Graph

Input: `Organizational data`: a set of annual reports of Standard & Poor Financial company documents in textual format .

`Model`: a pipeline to process unstructured text into a knowledge graph.

Output: $\mathcal{K}_G$ — A knowledge graph, based on stakeholder taxonomy classification, obtained from the extracted causal statements.

1: **for** each of the test documents in pdf form uploaded by the user. **do**
2: Extract the text from the pdf document.
3: Classify whether a sentence is causal or not using a transformer-based deep learning model.
4: Extract the causalities from the classified causal sentences.
5: Classify the extracted causalities based on the stakeholder taxonomy.
6: Construct the Knowledge graph $\mathcal{K}_G$
7: **end for**
8: **return** $\mathcal{K}_G$

The results of this causality extraction task were good compared to many baseline similar works [70, 149]. Since the results are promising for the causality extraction task, we decided to use organizational data as the training data for the transfer learning task.

5.2 Causality extraction on the medical data

This section discusses the data preparation for the medical data. It also discusses causal sentence classification and causality extraction on the annotated corpus of the medical text.

5.2.1 Data preparation and preprocessing

Seven documents on gestational diabetes guidelines provided by different societies are downloaded as PDF documents. The PDFs are converted into a document format,

and the documents are given to the annotators for annotating them manually. The annotators used tags to annotate the documents.

After annotating them, the NLTK sentence tokenizer is used to extract sentences from all the documents. The sentences from all the documents are appended together and converted into a data frame. Regular expressions are used to extract the causal sentence. If any of the sentences contain a tag <>, it will be extracted as a causal sentence. Again regular expressions are used to extract the phrases of cause, effect, action, signal, and condition from the sentences. The extracted phrases are used for computing inter-annotator agreement.

5.2.2 Causal sentence classification from medical data

For the causal sentence classification, we consider the manually annotated causal sentences as positive cases and label them as 1. An equal number of sentences from the corpus that are not causal are considered negative samples and are annotated as 0. The negative sentences are randomly selected from the list of the non-causal sentences. The data for this causal sentence classification consists of an equal number of causal and non-causal sentences. We ran four models on this data, out of which BERT gave a higher macro average F1-score of 0.94. The dataset is split into train and test using a five-fold cross-validation. Then BERT for sequence classification was fine-tuned on our dataset for four epochs. The results of this classification task for the highest performing model are given in Table 5.2

We ran Logistic regression, BiLSTM, and DistilBERT on the same data. With Logistic regression, a macro average F1-score of 0.85 was obtained, BiLSTM gave an average F1-score of 0.83, DistilBERT gave an average F1-score of 0.97, four shot prompting on GPT-4 gave an average F1-score of 0.80 which is lower than all other models.

A summary of results for all the models is given in Table 5.3. From the results, we can infer that fine-tuning BERT gives much better results, which is also the scenario

Table 5.2: Results of causal sentence classification on the medical data using BERT. Here the positive samples are the manually annotated data, and an equal number of negative samples are taken from the list of non-causal sentences from the medical data. With a five-fold cross-validation, BERT got a higher F1 score compared to all other models.

	Precision	Recall	F1-score	Support
Non-Causal	0.97	0.90	0.93	292
Causal	0.91	0.97	0.94	292
Accuracy			0.94	584
Macro average	0.94	0.94	0.94	584
Weighted average	0.94	0.94	0.94	584

Table 5.3: Summary of the results of the causal sentence classification on the medical data using various models. Here the manually annotated samples are considered as positive samples, and an equal number of data that are non-causal are selected as negative samples. From the results, we can understand that BERT performs better than all other models with a higher F1-score of 0.94. With GPT-4, a four-shot prompting was done.

	Precision	Recall	F1-score	Accuracy
Logistic regression	0.87	0.85	0.85	0.85
Bi-LSTM	0.91	0.76	0.83	0.85
DistilBERT	0.93	0.93	0.93	0.93
BERT	0.94	0.94	0.94	0.94
GPT-4 prompting	0.82	0.81	0.80	0.81

with the organizational data.

The detailed results of logistic regression and DistilBERT are given in Appendix A Table A.3 and Table A.4

We chose a four-shot prompting for the sequence classification based on our set of experiments for causal sentence classification on the Causal News Corpus dataset. The summary of the results of sequence classification with various prompt sizes on the Causal News Corpus is given in Appendix A Table A.5. From Table A.5 we can understand that there is no improvement in the performance with an increase in the prompts.

In real-world data, the distribution of positive and negative samples may not be the same. In gestational diabetes, there was a high imbalance between the causal and

non-causal sentences. There was a total of 290 causal sentences and 1900 non-causal sentences. In order to estimate the performance of the model on such real-world data, we appended the imbalanced causal and non-causal sentences and tested them using BERT. With BERT, we got an F1 score of 0.51 for the causal sentence classification. The results indicate that if one of the categories has more examples than the other and the data is highly imbalanced, the performance is high for the category with more examples than the other category. Also, the overall performance decreased. These results suggest that training should be done with oversampling the causal/non-causal data, or an equal amount of causal and non-causal sentences should be selected.

5.2.3 Causality extraction from medical data

This section elaborates on the approaches employed for causality extraction and the resulting outcomes. Given the good performance of DistilBERT with organizational data, this model was also applied to the medical data. Considering the limited sample size in medical data, we attempted to improve the learning process by increasing the number of epochs. This approach allows for more refined fine-tuning of the model.

In order to decide on the correct number of epochs and to avoid overfitting, we tried running the model for 100 epochs and plotted the validation loss and the training loss. The graph showing the train and validation loss for our highest performing model, BioBERT, is given in Figure 5.1.The graph for the DistilBERT and BERT are given in Appendix B Figure B.1 and Figure B.2

From the graph, we can understand that with the increase in the number of epochs, the training loss is constantly increasing and approaching 0. The validation loss decreases till 18 epochs and then starts to increase. Based on this, we fine-tuned DistilBERT for 18 epochs, BERT(BERT-base-uncased) for 20 epochs, and BioBERT for 16 epochs.

The data is split into train and test. DistilBERT for token classification is fine-tuned on the training data for 18 epochs. On the test data, the model obtained an

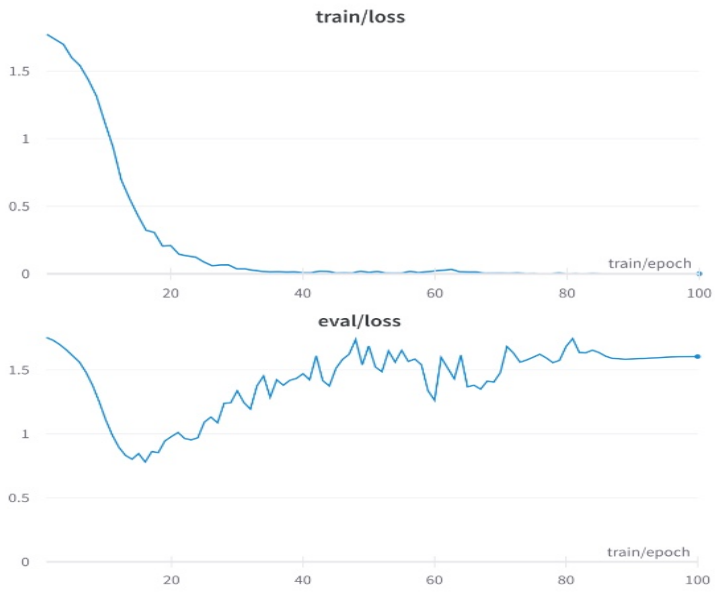


Figure 5.1: Graph showing the train and validation loss when fine-tuning on BioBERT. Looking at the graph, we can understand that with the increase in the number of epochs, the training loss is constantly decreasing and approaching 0. The validation loss decreases till 16 epochs and then starts to increase. Based on this, we fine-tuned BioBERT for 16 epochs.

average F1-score of 0.57.

Similarly, we fine-tuned BioBERT for 16 epochs and BERT for 20 epochs. Out of these three models, BioBERT[17] gave us an average higher F1-score. BioBERT gave an average F1 score of 0.61, and BERT gave an average F1 score of 0.60. The detailed results of fine-tuning BioBERT on the test data are given in Table 5.4. The detailed results for the DistilBERT and BERT are available in Appendix B Table B.2 and Table B.3.

From the detailed results, we can understand that the performance of the signal is 0 with all three models (BioBERT, DistilBERT, and BERT). In the overall dataset, there is only a small number of samples of the signal (Figure 4.2), and the model cannot predict it correctly. The higher the number of samples in the dataset, the performance of that particular category is better.

Table 5.4: Causality extraction results on the medical data using BioBERT, the highest performing model. Each token in the text was assigned a label Signal(S), Effect(E), Other(O), Cause(C), Condition(CO), and Action(A). The results are obtained by splitting the manually annotated data into train and test data. Note: there are very few samples in the entire dataset for the signal, and the F1-score for a signal is 0 with all the models.

	Precision	Recall	F1-score	Support
S	1.00	0.00	0.00	31
E	0.82	0.75	0.78	696
O	0.78	0.74	0.76	1186
C	0.62	0.69	0.65	411
CO	0.80	0.63	0.71	717
A	0.65	0.85	0.73	838
Accuracy			0.73	3879
Macro average	0.78	0.61	0.61	3879
Weighted average	0.75	0.73	0.73	3879

Table 5.5: Summary of the results of causality extraction on medical text using the Pre-trained Language Model (BERT) and its variants. The gestational diabetes data is split into train and test data. All the models are fine-tuned on train data and tested on test data.

	Precision	Recall	F1-score	Accuracy
DistilBERT	0.74	0.57	0.57	0.69
BERT	0.77	0.60	0.60	0.73
BioBERT	0.78	0.61	0.61	0.73

5.2.3.1 GPT-4 for causality extraction from medical guidelines

Apart from using one of the state-of-the-art performing transformer-based models (BERT), we ran two Large Language Models (LLMs), GPT-4 and LLAMA2, on the medical guidelines data for the causality extraction task. We did a few-shot prompting on this data with different prompt sizes and documented our results in this section.

Generative Pre-trained Transformer 4 (GPT-4) [13] outperforms most of the state-of-the-art performing models on the traditional NLP benchmark datasets. We explored various prompt sizes (zero, four, six, eight, ten-shot, and twenty-shot prompting). When token-level labels are given as prompting examples, the model hallucinated by predicting a long number of "O-other" tokens.

Note: For the organizational data, we did a few-shot prompting using GPT-3.5, which we performed as a token classification. But GPT-4 cannot deal with token classification. For the causality extraction, using medical data, we did a phrase-level extraction using both GPT-4.

Initially, with GPT-4, we did a four-shot prompting. The annotated data with the tags will be given as an example in the prompt, and the model is expected to predict similarly. A sample is given in Example 5.2.1.

Example 5.2.1 *<C>Gestational diabetes</C> has also been associated with an <E>increased risk of several long-term health outcomes in pregnant persons and intermediate outcomes in their offspring</E>*

We tried converting the predictions with the tags into a token-level format in order to compute the F1 score. However, since the tags are placed in different places in some of the gold annotations and predictions, the number of tokens in gold and predictions doesn't match. An example is given below

Example 5.2.2 ***Gold:*** *Importance<C>Gestational diabetes</C> is diabetes that develops during pregnancy.1-3 Prevalence of gestational diabetes in the US has been*

estimated at 5.8% to 9.2%, based on traditional diagnostic criteria, although it may be higher if more inclusive criteria are used.4-8 <C>Pregnant persons with gestational diabetes</C> <E>increased risk for maternal and fetal complications, including preeclampsia, fetal macrosomia (which can cause shoulder dystocia and birth injury), and neonatal hypoglycemia</E> .3,9-11 <C>Gestational diabetes</C> has also been associated with an <E>increased risk of several long-term health outcomes in pregnant persons and intermediate outcomes in their offspring</E> .12-16Table 1.

Prediction: *Importance Gestational diabetes is diabetes that develops during pregnancy. 1-3 Prevalence of gestational diabetes in the US has been estimated at 5.8% to 9.2%, based on traditional diagnostic criteria, although it may be higher if more inclusive criteria are used.4-8 <C>Pregnant persons with gestational diabetes</C> <E>increased risk for maternal and fetal complications, including preeclampsia, fetal macrosomia (which can cause shoulder dystocia and birth injury), and neonatal hypoglycemia. 3,9-11</E> <C>Gestational diabetes</C> has also been associated with an <E>increased risk of several long-term health outcomes in pregnant persons and intermediate outcomes in their offspring.12-16Table 1.</E>*

In Example 5.2.2, the phrases marked indicate the scenario where some extra spaces can be added, leading to the indifference in the number of tokens between gold and the predictions. In the gold data, `neonatal hypoglycemia</E> .3,9-11 have a space after the tag, but in the prediction, the tag is predicted after the number, which leads to no space between </E> and .3,9-11.` In some scenarios, the GPT-4 omits some of the words if they do not contain a causal relation (omits the 'O' labels in some places). This mismatch between the gold and the predictions impedes the token-level comparison and reporting of the F1 score. An example is given below:

Example 5.2.3 ***Gold:****Race/Ethnicity/Hemoglobinopathies<C>Hemoglobin variants </C> can <E>interfere with the measurement of A1C</E>, although most assays in use in the U.S. are unaffected by the most common variants.*

Prediction: *<C>Race/Ethnicity/Hemoglobinopathies variants</C> can interfere with the measurement of A1C, although most assays in use in the U.S. are unaffected by the most common variants.*

In Example 5.2.3, in the prediction, the keyword "Hemoglobin" is missing, which is present in the gold data. In some places, such inconsistencies lead to token mismatch between the gold and predicted data.

To compare the performance of GPT-4 with other models, the predictions are converted into the token level and manually checked to convert both the gold predictions to the same number of tokens for the four-shot prompting. In the predictions, some tokens are missed; those tokens are added to the predictions and marked as label "O."(as O indicates tokens that are not cause, effect, condition, action, or signal). After converting the data into a token level, we computed the F1 score. With GPT-4, we got an average F1 score of 0.32 with four-shot prompting. The detailed results of four-shot prompting with GPT-4 are given in Appendix B Table B.4.

As the predictions of GPT-4 are unpredictable and miss some tokens in a sentence leading to token mismatch, Jaccard distance and Cosine similarity are proposed as alternative solutions. The Jaccard similarity was computed using the textdistance[3] Python library. The cosine similarity is obtained by computing the vectors of both the gold and the predictions using the Universal Sentence Encoder[150]. The computed values are used to compute the pairwise cosine similarity between two vectors using Scikit-learn [4]. The cause, effect, signal, condition, and action are extracted from the predictions using regular expressions on the tags. The extracted prediction phrases

[3] `https://pypi.org/project/textdistance/`

[4] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html`

Table 5.6: The phrase level comparison results of few-shot prompting using GPT-4. We tried various prompt sizes (zero, four, six, eight, ten, and twenty-shot prompting). From the results, we can understand that the Jaccard similarity at ten-shot prompting is higher (higher the similarity, higher overlap between the gold and predicted spans), cosine similarity is lower (lower the similarity, higher the gold and press are related), and the F1-score between the labels is higher, after which the similarity and F1 decreases at twenty-shot. The cosine similarity, which gives the semantic similarity between gold and predictions, remains the same with all the prompt sizes. Here the F1-score is computed by comparing the gold labels and the predicted labels.

	Jaccard similarity	Cosine similarity	F1 (labels)
Zero-shot	0.37	0.20	0.27
Four-shot	0.41	0.20	0.35
Six-shot	0.55	0.20	0.47
Eight-shot	0.49	0.21	0.48
Ten-shot	0.55	0.19	0.50
Twenty-shot	0.55	0.21	0.28

and the gold annotated phrases are merged together. We perform three types of matching on the gold and predicted phrases.

- *Jaccard distance:* To measure the dissimilarity between the gold data and the predictions.

- *Cosine similarity:* To measure the semantic similarity between the gold data and the predictions.

The results of phrase level similarity between the gold annotated data and predictions of GPT-4 using various prompt sizes are summarized in Table 5.6

From the results of the various prompt sizes for the causality extraction on medical data, we can understand that the result of the ten-shot prompting gives a higher similarity and F1 score. There is not much difference in the cosine similarity with various prompt sizes. The Jaccard similarity gives the similarity score based on the overlap between the gold and the predictions. The higher the similarity, the more the similarity. Cosine similarity gives the semantic similarity between the gold and predicted phrases. The F1-scores are computed by comparing the gold labels with

Table 5.7: Summary of the results of the predictions of ten-shot prompting. Here the F1-score is computed by comparing the gold labels and the predicted labels. The F1 score for cause, effect, and action is higher compared to the other two labels.

	Precision	Recall	F1 score
Action	0.56	0.90	0.69
Cause	0.60	0.74	0.66
Condition	0.95	0.17	0.29
Effect	0.74	0.79	0.76
Signal	0.50	0.06	0.10
Accuracy			0.64
Macro average	0.67	0.53	0.50
weighted average	0.70	0.64	0.59

the predicted labels (Jaccard and cosine similarity for the predicted phrases, F1-score for the labels).

A recent work [151] uses ChatGPT on 13 benchmark datasets which include the datasets on causal relation and reasoning. In their experiment, they used zero-shot prompting along with in-context learning (few-shot learning) on three datasets: Choice of Plausible Alternatives (COPA) [152], e-CARE [153], HeadlineCause [154]. Their results indicate that, for reasoning and detecting causal relationships, ChatGPT exhibits strong performance. Their performance is particularly good with cause and effect. Our results indicate the same for the cause-and-effect scenario. The detailed F1-score for the ten-shot prompting label match between gold and predictions is given in Table 5.7

From the results in Table 5.7, we can understand that the F1 score for cause, effect, and action is higher compared to the other labels.

5.2.3.2 LLAMA2 for causality extraction from medical guidelines

LLAMA2[14] is a pre-trained and fine-tuned Large Language Model. Three variants of LLAMA2 are available, which differ in the parameters. 7B, 13B, and 70B parameters are publicly available. LLAMA2 is trained on two trillion tokens of data. In our experiments, the LLAMA2 7B parameter is fine-tuned on the medical data. It

is fine-tuned using the HuggingFace autotrain (refer to Appendix B for the autotrain command used).

To fine-tune LLAMA2, the first step is to prepare the data. At first, when the model was fine-tuned and tested on the token level as BERT, LLAMA2 was predicting a long number of "O-other" as GPT-4. So we dealt with this as a phrase-level extraction problem. The data is prepared with three parts which are instruction, input, and output. A sample training data is given in example 5.2.3

Example 5.2.4 *###Instruction: Extract the cause, condition, effect, signal, and action from the given sentence. ###Input: Pregnant persons with gestational diabetes are at increased risk for maternal and fetal complications, including preeclampsia, fetal macrosomia (which can cause shoulder dystocia and birth injury), and neonatal hypoglycemia. ###Output: ['Pregnant persons-signal', 'with gestational diabetes -cause', 'increased risk for maternal and fetal complications, including preeclampsia, fetal macrosomia (which can cause shoulder dystocia and birth injury), and neonatal hypoglycemia-effect']*

The test data should be similar to the training data except for the output, which should be empty. The gestational diabetes annotated data was split into train and test data. The HuggingFace autotrain [5] for the LLM fine-tuning was used to fine-tune the model. The fine-tuned weights are pushed into the HuggingFace dataset for inference. This experiment was done using Google Colab Pro+ with a High-RAM A100 GPU. Similar to the GPT-4, the predictions of LLAMA-2 were also at phrase level. So a similar evaluation strategy is followed for LLAMA2. We present the results with three types of distance.

The predictions are split into phrase levels and then compared with gold data. The Jaccard similarity was computed using the textdistance[6] Python library. The cosine

[5] `https://huggingface.co/docs/autotrain/llm_finetuning`
[6] `https://pypi.org/project/textdistance/`

similarity is obtained by computing the vectors of both the gold and the predictions using the Universal sentence encoder[150]. The computed values are used to compute the pairwise cosine similarity between two vectors using Scikit-learn [7].

Initially, we split the data into train and test using the Scikit learn train_test_split(). We have converted the phrase-level predictions into token-level. In the test data, there were a total of 59 samples. Out of the 59 samples, only 29 samples, LLAMA2 predicted the labels, so the evaluation is only for those sentences. With LLAMA2, we got an average F1-score of 0.36, which is lower than that of all the other models. The detailed results are available in Appendix B Table B.5

Since the test data size is very small, we have also tried a four-fold cross-validation on this data. The results of fine-tuning LLAMA2 using four-fold cross-validation with 3,5, and 10 epochs are given in Table 5.8.

With the increase in the number of epochs, the Jaccard similarity increased till ten epochs, but the F1-score remains the same after five epochs. Also, the predictions of LLAMA2 missed labels in many of the predictions. It extracted the phrases with no label. With three epochs, LLAMA2 missed 38% of the labels; with five epochs, 21% of the labels; and with ten epochs, it missed 26% of the labels. We omitted the predictions with no labels (108 predictions, 60 predictions, 76 predictions). The results of causality extraction presented in Table 5.8 are after omitting the predictions with no labels.

From the results, we can understand that Jaccard similarity and cosine similarity increase with the increase in the number of epochs. However, the F1-score remains the same with the increase in the number of epochs beyond 5 epochs. Also, the number of missed labels started increasing after 5 epochs.

[7] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html`

Table 5.8: The phrase level comparison results of LLAMA2 using 4 fold cross validation. Jaccard similarity and cosine similarity indicate the average similarity between the gold and the predictions. The F1 score is the comparison between the gold labels and predicted labels.

	Jaccard similarity	Cosine similarity	F1-score
LLAMA2 (3epochs)	0.70	0.17	0.62
LLAMA2 (5epochs)	0.87	0.18	0.71
LLAMA2 (10epochs)	0.91	0.19	0.71

5.3 Modals and Degree of Influence

Modal verbs such as 'can', 'could', 'may', 'might', 'will', 'would', 'shall', 'should', 'must', and 'ought to' significantly contribute to linguistic expression in English [155]. They indicate various modes such as necessity, potential, permission, and obligation by modifying the main verb in a sentence, providing additional context that aids in conveying the speaker's intention. Thus, proper comprehension and application of modal verbs are vital for effective English communication.

In medical texts, modal verbs are crucial as they offer essential context and specificity. They can signify certainty, potential, permission, or requirement, all of which are essential in medical communication. For instance, 'must' can convey urgency or a medical necessity, as in "Patients must take this medication daily." 'May' or 'might' can illustrate potential or uncertainty, often used when discussing potential side effects or outcomes, such as "Patients may experience dizziness." 'Should' is often employed to provide medical advice or recommendations, like "Patients should avoid heavy lifting after surgery." Therefore, accurately understanding and interpreting modal verbs is significant for precise communication in medical texts.

A few years ago, a work [156] on protein-protein extraction from medical texts explained the importance of understanding the modal verb during the extraction process. A relatively recent work [157] proposes an event extraction system that can capture various types of modality.

In this study, modal verbs are automatically extracted from the annotated corpus,

Table 5.9: Summary of the results of extracting modal verbs from the sentences. All three models predict the modals correctly. In these experiments, we did a two-fold cross-validation.

	Precision	Recall	F1-Score
DistilBERT	0.95	0.99	0.97
BERT	0.96	0.97	0.97
BioBERT	0.96	0.99	0.97

supporting the identification of the intensity of the recommendation or assertion. Annotators were given a list of English modal verbs[8], which they used to annotate sentences containing modal verbs. Additionally, sentences containing modals were automatically extracted using regular expressions. In total, 128 causal sentences contained modal verbs. The transformer-based models (BERT, DistilBERT, BioBERT) were fine-tuned using the 128 sentences that contain modal verbs. This was done by dividing the data into training and testing datasets. Owing to the small number of samples, the model's performance was suboptimal. To enhance the performance, we attempted to increase the sample size by including examples from a similar dataset. We randomly selected samples from a comparable dataset [110], ending up with a total of 1023 sentences. Each sentence with a modal verb was annotated at the token level, marking the modal verb as MD and all other tokens as O. Every token in the testing data was then categorized as MD or O. The results of these modal predictions are provided in Table 5.9.

From the results, we can understand that all three models predict the modal verbs correctly. We also looked into the vocabulary of all three models, which contains almost all the modal verbs. With this experiment, we can understand that modal verb extraction can be done easily with BERT. If we can identify such modal verbs, we can employ some modifications to BERT to understand the modal sense [158]

[8] `https://en.wikipedia.org/wiki/Modal_verb`

Table 5.10: Summary of the results of classifying the sentences with degree of influence. Logistic regression is performing better. GPT-4 is also performing almost similarly to the best-performing model.

	Precision	Recall	F1-Score	Accuracy
DistilBERT	0.66	0.63	0.62	0.63
BERT	0.77	0.76	0.76	0.76
Logistic regression	0.94	0.94	0.94	0.94
GPT-4	0.93	0.93	0.92	0.93

5.3.1 Degree of influence

The degree of influence indicates the impact a cause has on its effect. Clinical Practice Guidelines (CPGs) are documents designed to assist physicians in their decision-making process. These are written in a manner that specifically outlines the cause and its effect. In certain instances, these guidelines specify the extent to which a cause influences its effect.

Example 5.3.1 *If a patient smokes more than 20 cigarettes per day and more than five drinks a day have a hazard ratio of 8.5 [18, 19]*

Example 5.3.1 demonstrates that an individual who smokes 20 cigarettes a day and consumes over five alcoholic drinks daily has an elevated likelihood of developing lung cancer. The hazard ratio in this scenario determines the probability of a person in a higher-risk category reaching the endpoint (developing lung cancer in this case) prior to those in lower-risk groups. Recognizing such information is beneficial in the decision-making process. The outcomes of identifying the degree of influence are provided in Table. 5.10.

Identifying the degree of influence is considered a classification task. The sentences with the degree of influence are labeled as positive classes, and an equal number of sentences that do not have a degree of influence are labeled as negative classes.

The results presented in Table 5.10 indicate that the model can identify the sentences with a degree of influence even if it is fine-tuned on small data. In the entire

dataset, there were only 42 sentences with a degree of influence that indicated a relation between the cause and effect. If more data is available, this classification task can be extended to extract the degree of influence from the sentence. If extracted, the degree of influence can be used to find how much influence a cause has on effect, which can be used for the decision-making process.

5.4 Inter-sentence level causal sentence classification

Identifying causal sentences from the entire text document is one of the important steps in the causality extraction task. In recent times, various works on causal sentence classification have been published [159, 160, 161]. However, all the work concentrates on the intra-sentence level classification. All the work on the causality extraction at the inter-sentence level concentrates only on the causality extraction [162, 163, 164]. A recent work [165] generates a word-level encoding and uses an attention-based Bi-LSTM to compute the weight of all words. Then a graph attention network is introduced to get the information in the neighboring nodes, which is then sent to the sentence-level BiLSTM. Another similar work [166] proposes a document-level Event Causality Identification that detects the causal relation between the event mention pairs in text. To the best of our knowledge(per Google Scholar search)[9], there is no work that does a cross-domain inter-sentence level causal sentence classification on a small dataset. This work considers only the cause/effect relationship in two consecutive sentences because, in both the organizational and medical data, the cause/effect relationships are annotated between two consecutive sentences. From the classified inter-sentence level causal sentence classification, causalities can be extracted.

Inter-sentence level causal sentence classification is performed on two datasets, organizational data and medical data. For the organizational data, when the cause/effect relationship is present across the sentence, they are combined together as a single row

[9]Search query - Cross domain inter-sentence causal sentence classification

Table 5.11: Summary of the results of inter-sentence causal sentence classification on the organizational data. Bi-LSTM is performing better with the inter-sentence level classification compared to other Pre-trained Language models like BERT. The F1-score increased till three layers were stacked. On adding the fourth layer, the F1-score decreased.

	Precision	Recall	F1-Score	Accuracy
DistilBERT	0.51	0.51	0.51	0.51
BERT	0.52	0.52	0.52	0.52
Logistic regression	0.52	0.52	0.52	0.52
Bi-LSTM	0.85	0.84	0.85	0.85
Stacked Bi-LSTM (2 layers)	0.87	0.87	0.87	0.87
Stacked Bi-LSTM (3 layers)	0.89	0.89	0.89	0.89

which will be considered as positive samples. An equal number of negative samples are selected. The negative samples are selected based on two scenarios.

- First, the sentences that are not causal are selected from the annotated set of documents. Two consecutive sentences from the corpus are appended together to match the positive samples.

- Second, an equal number of sentences are generated by data augmentation using Chat GPT-4. Similar sentences are generated based on the sentences that are selected from the corpus.

Both the positive and negative samples are shuffled and split into train and test data. The results of the inter-sentence causal sentence classification are summarized in Table 5.11

Totally there were 641 inter-sentence relations in the organizational data. We appended an equal number of the sentences that are non-causal. So there are totally 1282 sentences in the data that were used for the classification. With organizational data for the BERT, DistilBERT, we fine-tune the model for four epochs that are recommended by the original BERT authors. We use a 5-fold cross-validation.

In the medical data, there were only 38 sentences that had inter-sentence relations. So we did a cross-domain inter-sentence classification with the medical data. The

Table 5.12: Summary of the results of inter-sentence causal sentence classification on the medical data. A cross-domain inter-sentence level causal sentence classification is done. Here the model is fine-tuned on the inter-sentence level organizational data and tested on the inter-sentence level medical data.

	Precision	Recall	F1-Score	Accuracy
DistilBERT	0.76	0.53	0.39	0.53
BERT	0.75	0.50	0.33	0.50
Logistic regression	0.38	0.43	0.37	0.43
Bi-LSTM	0.48	0.49	0.44	0.49
Stacked Bi-LSTM (2 layers)	0.55	0.54	0.51	0.54
Stacked Bi-LSTM (3 layers)	0.41	0.42	0.40	0.42

inter-sentence relations in organizational data are used as training data, and the medical data is used as the test data. From the results in Table 5.11 and Table 5.12, we can understand that the Bi-LSTM can capture the relation between the sentences. With a three-stack Bi-LSTM, an average F1-score of 0.89 is obtained. With the cross-domain classification, Bi-LSTM with two layers gave a higher F1-score of 0.51. This can be improved by appending some similar medical datasets with inter-sentence relations like [167].

In this Chapter, we presented the results of the causality extraction and causal sentence classification on both organizational data and medical data. We have also shown the importance of the modal verbs in medical text and the degree of influence. Lastly, we have shown the possibility of an inter-sentence causal sentence classification. The performance of the causality extraction and degree of influence can be improved by increasing the size of the data. We have shown that the model can extract the modal verbs from the text successfully. We have established a baseline for the cross-domain inter-sentence causal sentence classification. The applicability of using LLMs for the causality extraction on medical guidelines was shown, and we are working on improving its performance by applying cross-validation and increasing the prompt size.

CHAPTER 6: CAUSAL TRANSFER LEARNING FROM ORGANIZATIONAL TEXT TO MEDICAL TEXT

Machine learning is used widely for many applications where the machine learning model will be trained on certain data in order to predict future outcomes. Usually, the machine learning model will be trained and tested on a dataset that will have the same data distributions. If there is a difference in the data distributions between the train and test datasets, the performance of the model can be degraded [168]. However, it will be a difficult task to create an annotated dataset for all domains. The motivation behind transfer learning is to have high-performance predictions on the target domain when the model is trained on a related source domain.

Transfer learning, or domain adaptation, ([169], [170]) have been proposed as a mitigation for the problem of scarcity of annotated data. The idea is that the performance of a machine learning program can be enhanced by pretraining on a related task. A survey article [171] defines the transfer learning as follows:

> "Given a source domain $\mathtt{D}_S$ with a corresponding task $\mathtt{T}_S$ and a target domain $\mathtt{D}_T$ with corresponding task $\mathtt{T}_T$, where $\mathtt{D}_S \neq \mathtt{D}_T$ or $\mathtt{T}_S \neq \mathtt{T}_T$, transfer learning aims to improve the performance of the model's predictions by using the related information from $\mathtt{D}_S$ and $\mathtt{T}_S$."

Transfer learning problems can generally be categorized into two main classes based on how similar the source and the target domains are. The main classes are Homogeneous transfer learning and Heterogeneous transfer learning. In the homogeneous transfer learning, the source ($\mathtt{X}_s$,$\mathtt{Y}_s$) and the target domain($\mathtt{X}_t$,$\mathtt{Y}_t$) will be similar ($\mathtt{X}_s$=$\mathtt{X}_t$ and $\mathtt{Y}_s$=$\mathtt{Y}_t$), and the task will be to bridge the gap between the source and the target

domains. In heterogeneous transfer learning, the source and the target have different feature spaces ($\mathtt{X}_s \neq \mathtt{X}_t$ and $\mathtt{Y}_s \neq \mathtt{Y}_t$).

To solve the transfer learning problem, different strategies and implementations are employed. Three general strategies that are used to solve homogeneous transfer learning are a) the marginal distribution difference in the source should be corrected [172],[173], b) correct the conditional distribution difference [174],[175], or c) correct both the marginal and the conditional distribution differences [176],[177]. With heterogeneous transfer learning [178], [179], the source and the target domains will be aligned with the assumption that they are equal. If the distributions are not equal further domain adaptation will be needed. Transfer learning does not always produce positive results [180]

In NLP, transfer learning has been used in the Natural Language Inference (NLI) task (e.g. [181]), and for various other tasks like causal sentence detection ([72]); finding condition-action sentences in medical guidelines ([182]) and understanding of biomedical texts [183]. In another example, extracting drug timelines from the electronic health record was done by [184] by training the model on THYME colon cancer corpus and testing on THYME brain cancer corpus.

Transfer learning is often used interchangeably with the term *domain adaptation*, especially in natural language processing, as observed by [170].

Transfer learning focuses on applying the knowledge gained in solving a problem to a different but related problem [185], [186]. Because of the limited annotated data available for the medical guidelines text, the success in transfer learning will help expand the corpus of annotated data for the medical text. Transfer learning on medical guidelines has given promising results for [182], but this work tests the domain adaptability between different medical guidelines. They use rhinosinusitis and hypertension guidelines for training and asthma guidelines for testing the domain adaptability. This work will test domain adaptability between the financial data and

the medical guidelines text. Since there are a lot of annotated datasets, as listed in Table 3.1,2.1 from the various competitions are available, the medical guidelines annotated corpus can be expanded with success in transfer learning.

In this chapter, we discuss two sets of transfer learning experiments.

- Experiment 1 – Run a set of experiments to understand the relationship between properties of datasets and the degree of success of transfer learning.
- Experiment 2 – Apply Experiment1 results to document the transfer learning results between Organizational data and the medical data.

6.1 Understanding the relation between the similarity of data and degree of success of transfer learning

In this section, we investigate the applicability of transfer learning (domain adaptation) to address the impediments to the availability of annotated data in experiments with three publicly available datasets: FinCausal [95], SCITE [70], and Organizational [11]. We perform pairwise transfer experiments between the datasets using DistilBERT (a variant of BERT) and measure the performance of the resulting models. To understand the relationship between data sets and performance, we measure the differences between vocabulary distributions in the datasets using four methods: Kullback-Leibler (K-L) divergence, Wasserstein metric, Kolmogorov-Smirnov test, and Maximum Mean Discrepancy. We record the predictive values of each measure.

6.1.1 Measures of divergence and their uses

There are infinitely many ways we can talk about differences between text data. However, the simplest measures of difference are count-based, i.e., statistical. We use four popular tests for differences between the distributions: Kullback-Leibler divergence, Wasserstein distance, Kolmogorov-Smirnov test, and Maximum Mean Discrepancy. The mathematical relations between them are described in [187]. However, in

this work, we care about their potential predictive powers with respect to the accuracy of transfer learning for causality extraction. All three tests have been used in NLP, and KL divergence is perhaps the most popular.

KL divergence is a statistical distance that measures how different is a probability distribution compared to another. It is denoted by $D_{KL}(P \mid\mid Q)$ where P and Q are the probability distributions [188]. Notably, it is not symmetric $D_{KL}(P||Q) \neq D_{KL}(Q||P)$.

An example of recent use is shown in [189] to disentangle the syntax and semantics in a deep decomposeable model. For semantic similarity tasks and syntactic similarity tasks, their model improves their disentanglement quality.

In this work, using KL divergence, we show that the distributions of the three datasets are different and predict quite well transfer results.

KL Divergence measures the difference between two probability distributions based on information theory, that is, how much we can learn from one distribution about another, and therefore is not symmetric. The other two measures are symmetric. Wasserstein distance measures the distance between two probability distributions by considering the 'cost' of transforming one into the other and is symmetric. That's why it is sometimes called "the earth mover distance." Finally, the Kolmogorov-Smirnov test is a statistical test used to compare empirical distributions and is often employed to determine if a sample comes from a specific distribution and is not symmetric.

In this work, we are applying these concepts without any modifications. An introduction, formulas, and comparisons of mathematical properties are available in [187]. Examples of their uses in NLP appear e.g., in [188, 190] [191, 192] . For our practical objectives, we care about the existence of (Python) packages that we can easily apply to compute the required measures.

6.1.2 Datasets

In our experiments, we use three causality extraction datasets. First, SCITE [70], which extends the annotations of SemEval 2010 task 8 dataset [193] by considering all the causal triplets present in the sentence, whereas [193] considers only one causal triplet in the sentence. This dataset consists of text data from the web, which is not particularly related to the financial domain. Second, FinCausal[95], which is created as part of a challenge FinCausal 2022. This challenge aims to extract causalities from financial documents. This data is extracted from the 2019 financial news, which is collected from 14,000 economics and finance websites. Third, the Organizational (ORG) dataset, which was created for the causality extraction on Financial documents [11]. In this dataset, the 2020 SEC 10-K Documents of 65 S&P 80 Financial Companies were collected and manually annotated.

Here are some examples from each dataset. In the SCITE dataset, the cause-effect pairs are annotated using the XML tags, as shown below:

Example 6.1.1 *SCITE: <item id="15" label="Cause-Effect((e1,e2))"> <sentence> This case arises from <e1>;a December 21, 2005 automobile accident <e1> that resulted in <e2> the death <e2> of Larry Haynes.</sentence>*

In the FinCausal dataset, the cause-effect relation pairs are available as tags. <e1> represents the cause and <e2> represents the effect. The phrases of cause/effect are also available.

Example 6.1.2 *FinCausal:*

Text: It found that total U.S. healthcare spending would be about $3.9 trillion under Medicare for All in 2019, compared with about $3.8 trillion under the status quo. Part of the reason is that Medicare for All would offer generous benefits with no copays and deductibles, except limited
cost-sharing for certain medications.

Tag format: <e2>It found that total U.S. healthcare spending would be about $3.9 trillion under Medicare for All in 2019, compared with about $3.8 trillion under the status quo.</e2> <e1>Part of the reason is that Medicare for All would offer generous benefits with no copays and deductibles, except limited cost-sharing for certain medications.</e1>

The organizational data is annotated in the BIO-label format. For each of the tokens in the text, a label will be assigned. The cause is represented as C, effect as E.

Example 6.1.3 *Organizational:*

Text: ["When ","a ","policyholder ","or ","insured ","gets ","sick ","or ","hurt ","the ","Company ","pays ","cash ","benefits ","fairly ","and ", "promptly","for ","eligible ","claims"]

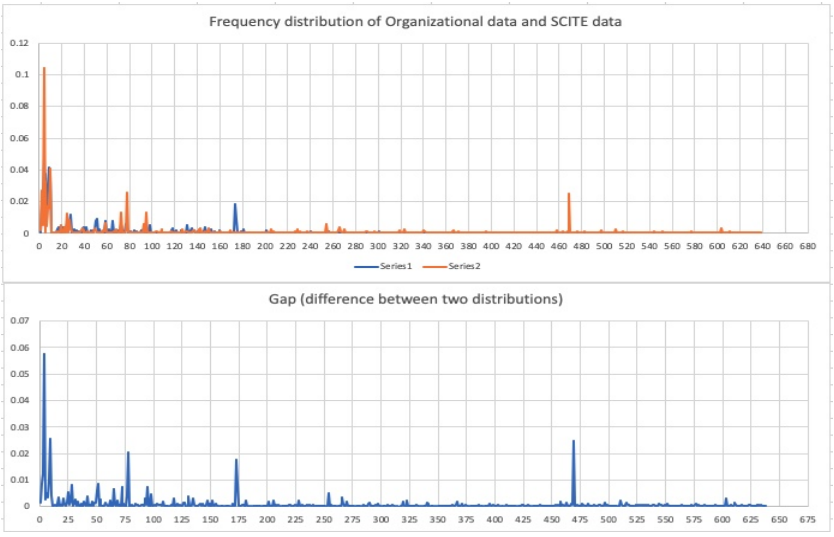
Label: ["O","O","B-C","I-C","I-C","I-C","I-C","I-C","I-C", "O", "O","B-CT", "B-E","I-E", "I-E","I-E","I-E","I-E","I-E","I-E"]

For the transfer learning experiments, we converted all these three datasets into the IO label format.

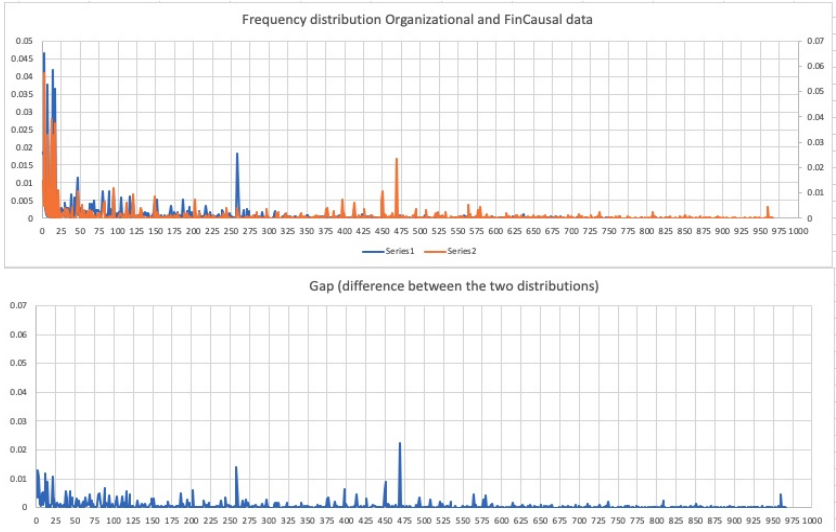
6.1.3 Differences between the datasets

To understand the differences in the distributions, we have created a feature distribution chart. This chart plots the frequency of the words in both the training and test data. The Organizational data (training data) had a total word count of 4747, and the SCITE data (test data) had a word of 1488. Totally 638 words were common in both of these datasets. Similarly, we have computed a frequency distribution chart for the Organizational and FinCausal dataset. In the FinCausal data, we had a total of 1595 words, out of which 966 are common in both datasets.

These differences in word distributions are shown in Figure 6.1 (and further quantified in Table 6.1).



(a) The top part of the chart with series1 and series2 indicates the frequency distribution on a 638 common word count between the Organizational and SCITE data. The bottom part, with only one blue legend, shows the gap between the two distributions at the top.



(b)The top part of the chart with series1 and series2 indicates the frequency distribution on a 966 common word count between the Organizational and FinCausal data. The bottom part, with only one blue legend, shows the gap between the two distributions at the top.

Figure 6.1: From the top and the bottom panel, we observe that the difference between the distributions is high for Organizational and SCITE, whereas the gap between the Organizational and FinCausal is smaller. As we shall see later, the difference in distribution is predictive of the F1-score in transfer learning. This is true both when we measure the differences by the K-L divergence and by the Wasserstein distance, although the former is more accurate.

Looking at graphs (a) and (b) of Figure 6.1, we can see the gaps between the pair's frequency distributions. The gap between FinCausal and Organizational data seems smaller than for the SCITE data.

Table 6.1: Summary of the computed K-L divergence values, Wasserstein distance (Wassers dist), Kolmogorov-Smirnov test (Kolmog-Smir), and Maximum Mean Discrepancy (MMD) on all the combinations of the datasets. The K-L divergence, MMD, and the Wasserstein distance on the same datasets is zero, meaning that there is a maximum overlap between the train and the test datasets. The higher the value of the K-L divergence, the lower the similarity between the datasets. For the K-S test, the low p-values prove that the distributions are different. From the computed values, we can understand that SCITE is less similar to FinCausal, and Organizational datasets.

Train data	Test data	K-L diverg	Wassers dist	Kolmog-Smir	MMD
SCITE	SCITE	0	0	1.0	0
	FinCausal	0.942	4.95	3.9555e-59	0.7796
	Organizational	0.906	13.1	1.1294e-136	0.7982
FinCausal	FinCausal	0	0	1.0	0
	SCITE	0.771	4.95	0.000109	0.1067
	Organizational	0.286	8.15	2.2483e-66	0.1488
Organizational	Organizational	0	0	1.0	0
	FinCausal	0.279	8.15	1.1639e-65	0.1481
	SCITE	0.336	13.1	3.3517e-151	0.3559

This is intuitively explained by the fact that the Organizational and the SCITE data are from completely different domains. Organizational data is created on the financial documents, whereas the SCITE is from the web text. Organizational and FinCausal seem to be similar data because they both are created using the financial text. But the Organizational data is annotated on the financial company reports, whereas FinCausal data is annotated on the financial text from the web.

As shown in Table 6.1, the K-L divergences between the data sets vary and confirm the impressions from Figure 6.1. Thus for the Organizational data and SCITE data we get the values 0.336 and 0.369; in contrast, for Organizational and FinCausal we get 0.279 and 0.286. In both cases and all directions the values are relatively high, which means the distributions are different.

We repeated the same set of comparisons using the Wasserstein metric, the Kolmogorov-Smirnov test and the Maximum Mean Discrepancy (MMD). The Wasserstein distance between the Organizational data and the SCITE data is 13.09, and the distance between the Organizational and FinCausal data is 8.14. We got a Wasserstein distance of 0 between the dataset with itself, and the distance between SCITE and FinCausal is 4.95.

The Kolmogorov-Smirnov (KS) test can be used to compare two probability distributions to check whether they are drawn from the same distribution. We chose the standard confidence level of 95%, which means the values that are in favor of the alternative will be rejected if the p-value is less than 0.05. All the p-values we obtained were much smaller than that, indicating and quantifying the differences between the word frequency distributions [1].

The Maximum Mean Discrepancy (MMD) is a non-parametric statistical test that can be used to identify the amount of discrepancy between the two probability distributions. It can be given as the distance between the feature means. The MMD between the same datasets will be 0. The maximum the MMD value, the maximum discrepancy between the source and target domain.

All the obtained values of differences between the data sets are summarized in Table 6.1. The table also suggests their intended experimental use in transfer learning. That is, training on one set and testing on another. The impact of these measured differences and the predictive value of each test is discussed in Section 6.1.4.

6.1.4 Experiments and Results

Given the performance of DistilBERT is good for the causality extraction task on all three datasets individually ([70], [95],[11]), the next natural question, namely is what happens when we attempt transfer learning, and if there are differences in

[1]For both computations, we used the SCIPY packages: `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html`, `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html#scipy.stats.ks_2samp`

performance.

To answer this question, we ran several transfer learning experiments with the three datasets. In all the experiments, the DistilBERT model was fine-tuned on one of the datasets, and the other two were used as the test data. For the FinCausal dataset and Organizational data, several BERT variants perform well, as reported in [194] and [11]. In another example, experiments using DistilBERT on the SCITE data produce a (relatively good) macro average F1-score of 0.88 (in our experiment). Here and later, the results are reported using macro-average scores; for example, F1 refers to the macro-average F1 score, i.e., the average F1 score for all the labels.

Before we discuss the results, we need to mention the composition of the datasets. Thus, SCITE contains the gold annotation for all train, validation, and test subsets. In contrast, in FinCausal, although we have split into train, validation, and test sets, there is no gold standard released for the test set, and therefore we use the validation set as test data. In Organizational (2235 sentences), we do the split train (70%), validation (10%), and test (20%).

We performed three sets of rounds of transfer experiments. In the first round of experiments, we fine-tuned the DistilBERT model on the SCITE train dataset, and we tested it on SCITE test data and on validation data of both FinCausal and Organizational test data. In the second round, we fine-tuned the FinCausal train data and tested on FinCausal validation and on Organizational and SCITE test data. And in the third set of experiments with transfer, we first fine-tuned the Organizational training data, and then tested on Organizational and SCITE test data and on FinCausal Validation data.

We are reporting our results of fine-tuning DistilBERT on the FinCausal data and predicting on the FinCausal, even though earlier results of [194] are available, showing the F1 of 0.87 on the validation data. Our results, for comparison, are 0.94, as shown in Table 6.2. The difference perhaps due to the fact that we use the Trainer() from

the Huggingface to fine-tune the model [2], whereas [194] use the transformer model from the Huggingface source [3].

The result of fine-tuning DistilBERT on the Organizational data and testing it on the Organizational data were obtained earlier and are presented in [11].

We train our model for (the optimal) 3 epochs with a batch size of 16. All the experiments are conducted on NVIDIA-SMI 525 GPU (using Google Colab). The results of running the causality extraction task on the SCITE, FinCausal, and Organizational dataset are summarized in Table 6.2.

Table 6.2: Summary of the transfer learning experiments. This table shows the performance of DistilBERT for causality extraction. In the top panel of the table, the model is fine-tuned on the Organizational data and tested on SCITE, the FinCausal dataset. In the bottom panel, the model is trained and tested on the FinCausal data.

Train data	Test data	P	R	F1
SCITE	SCITE	0.86	0.87	0.86
	FinCausal	0.8	0.04	0.075
	Organizational	0.68	0.07	0.11
FinCausal	FinCausal	0.94	0.94	0.94
	SCITE	0.21	0.74	0.33
	Organizational	0.64	0.75	0.69
Organizational	Organizational	0.78	0.89	0.83
	FinCausal	0.79	0.70	0.74
	SCITE	0.25	0.78	0.38

With respect to the transfer learning task, from Table 6.2, we can understand that the performance of the model is much better when the Organizational data is used as the training data and the validation data of FinCausal data is used as the test data, and vice versa. We got a macro average F1 score of 0.74 when FinCausal data was used as the test data and a macro average of 0.38 when SCITE data was used as the test data. There is almost a 0.36 increase in the F1-score when FinCausal is used as a test dataset rather than using SCITE. Similarly, we can see a 0.33 increase in the F1-score when FinCausal is used as a train, and the SCITE and Organizational data are used as test data.

[2] `https://huggingface.co/docs/transformers/tasks/token_classification`

[3] `gitclonehttps://github.com/huggingface/transformers.git`

It means that the model performs better when there is more similarity between the vocabulary used in the train and test dataset. From section 6.1.2, we know that the SCITE data is created from the web text, and the FinCausal data is annotated on the financial documents.

To understand how the F1-score varies depending on the K-L divergence, Wasserstein distance, and Kolmogorov-Smirnov test p-values, we have plotted the dependencies in Figure 6.2, in their simplest forms, as linear regression lines.

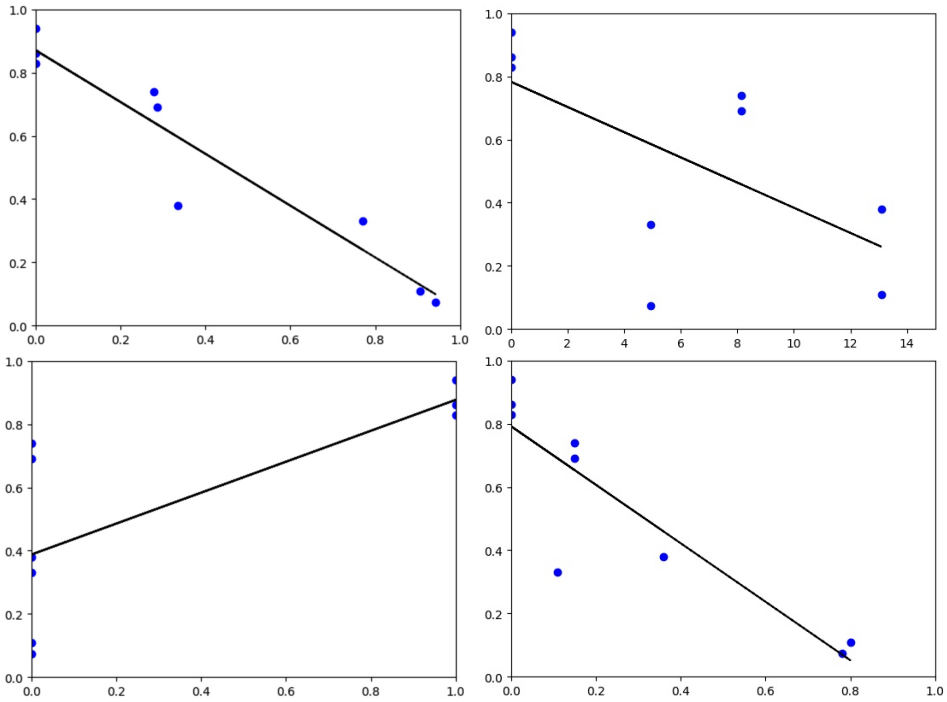


Figure 6.2: **Top Left Panel:** A linear regression model approximating between the data points of the experiments; the K-L divergence between data sets is represented along the X-axis, and the F1-scores of the corresponding machine learning transfer experiments are shown represented along the Y-axis. **Top Right Panel:** The results of the same experiments using the Wasserstein distance instead of the K-L divergence. **Bottom Left Panel:** The results of the same experiments using the Kolmogorov-Smirnov test. **Bottom Right Panel:** The results of the same experiments using the Maximum Mean Discrepancy (MMD).

We can see that lower K-L divergence values predict higher F1 scores. The same is true for Wasserstein distance and MMD. The dependence between the computed p-values of the Kolmogorov-Smirnov test and the F1 score is plotted in the bottom

left panel of Figure 6.2, but the diagram does not seem informative – the high F1 values correspond to identical distributions.

Linear regression helps quantify these impressions. Thus K-L divergence predicts the performance of transfer learning with high accuracy as measured by R2= [0.07462152] and confidence interval [-0.81916225 0.87093901]. The other three measures are not particularly good: surprisingly, Wasserstein distance gives R2 = [0.52912651], with the confidence interval [-0.03987017 0.78268854], the MMD gives R2 = [0.17625334], with the confidence interval [-0.92479103, 0.79202877], and the K-S test R2 = [0.40025979], with the confidence interval [0.48917194 0.38749281].

With a small number of points, we obtain wide confidence intervals. So, the results, even though confirming the observations of [182] on medical transfer learning, have to be taken with a grain of salt. Nevertheless, they do suggest the higher predictive value of K-L divergence for this and perhaps similar tasks.

We performed additional transfer experiments using SpanBERT [16]. The F1 scores are lower. Example: when fine-tuned on the Organizational data and tested on the SCITE's test data, SpanBERT gave a macro average precision of 0.25, recall of 0.74, and F1-score of 0.37; and when fine-tuned on Organizational data and tested on the validation data of the FinCausal dataset, SpanBERT gave a macro average precision of 0.80, recall of 0.64, and F1-score of 0.70. However, just like for DistilBERT, KL divergence is the best predictor of the successful transfer.

6.2 Transfer learning between Organizational data and Medical data

In the previous section, we selected three datasets from different areas with the aim of understanding the relationship between data similarity and the effectiveness of transfer learning. The results indicated that even with some resemblance between the training and test data, the transfer learning outcomes between FinCausal and Organizational data were much better than that of SCITE and FinCausal/Organizational. This is primarily due to FinCausal and Organizational data having more similarity

than SCITE. To validate this, we perform two series of experiments in this section. The first involves training the model on Organizational data and testing it on medical data, which is a case of cross-domain few-shot transfer learning. For the second experiment, we will also be implementing cross-domain few-shot transfer learning, where we randomly add 10% of the medical data to the Organizational data.

6.2.1 Cross domain zero-shot transfer learning

From Chapter 4, we can understand that organizational data was annotated on financial documents, and the medical data was annotated on a set of gestational diabetes Clinical Practice Guidelines. This indicates that both these datasets are annotated on completely different domains. In zero-shot transfer learning, we are fine-tuning the model on organizational data and testing the model with medical data. Since the organizational data has more samples than the medical data, we chose the organizational data as training data. A recent work on transfer learning for causality extraction task [195] pre-train the ELECTRA model on five of the available datasets(out of 5, two of them are annotated on medical text). They use these pre-trained weights for causality extraction on the text related to Sjogren's syndrome. The difference between this work and our work is that they are doing pretraining. The data includes the medical domain text. But we are just fine-tuning the model on an entirely different domain data. When preparing the dataset, the causal triggers in the organizational data are marked as the 'O-other' label. In the medical data, apart from cause and effect, there are various labels like signal, condition, and action. If all these labels are marked as 'O', the count of 'O' will be higher. Also, the sentences with condition and action alone will be considered non-causal. So we considered only the sentences with cause/effect relationship as the test data.

As a baseline model, we ran standard classifiers like random forest, logistic regression, and Naive Bayes classifier. Then we ran transformer-based models like BERT and its variants. We also ran the Large Language Model LLAMA2. The summary of

Table 6.3: Summary of the causal transfer learning results. Here the financial data is used for training, and the gestational diabetes guidelines are used as test data. Precision, Recall, and F1-score given are the macro average scores. From the results, we can infer that the F1 score on DistilBERT is higher compared to the other models.

	Precision	Recall	F1-Score	Accuracy
Random Forest	0.34	0.34	0.32	0.40
Multinomial Naive Bayes	0.38	0.35	0.21	0.22
Logistic regression	0.38	0.39	0.26	0.26
DistilBERT	0.51	0.51	0.49	0.51
BERT	0.50	0.52	0.48	0.48
BioBERT	0.59	0.63	0.59	0.60

Table 6.4: Summary of the causal transfer learning results when BioBERT was used. Here the financial data is used for training, and the gestational diabetes guidelines are used as test data. From the results, we can infer that the F1 score on BioBERT is higher compared to the other models.

	Precision	Recall	F1-Score
E	0.62	0.68	0.65
O	0.74	0.51	0.61
C	0.41	0.70	0.52
Accuracy			0.60
macro avg	0.59	0.63	0.59
weighted avg	0.64	0.60	0.60

zero-shot causal transfer learning for various models is given in Table 6.3

Looking into the results summarized in Table 6.3, we can understand that the transformer-based models are performing better than the baseline classifiers. As we did in Section 5.2.3, we ran all the models for 100 epochs to decide on the number of epochs and to avoid overfitting. With DistilBERT, the validation loss starts to increase after 4 epochs constantly. Similarly, for BioBERT after 8 epochs and for BERT after 7 epochs the validation loss increased constantly. A graph showing the train and the validation loss is given in Figure 6.3. The graph for the BERT and BioBERT are given in Appendix C, Figure C.1 and Figure C.2

The detailed results of the highest-performing model, BioBERT, are given in Table C.2. The detailed results of all other models are given in Appendix C

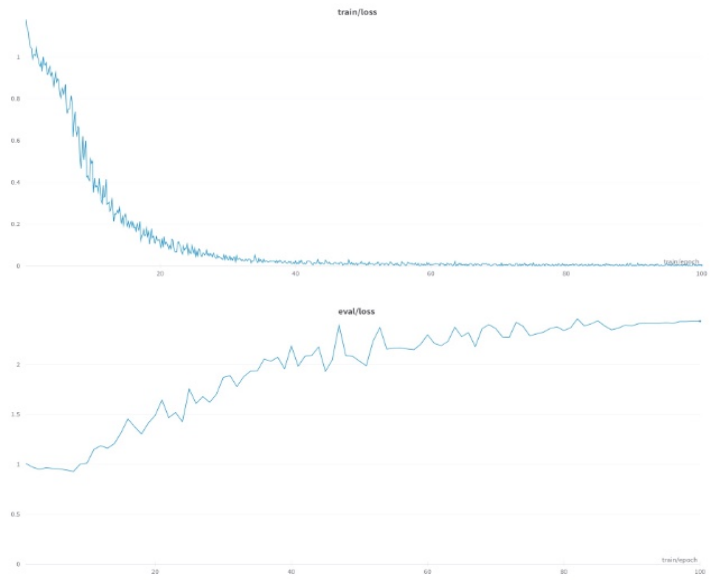


Figure 6.3: Graph showing the train and validation loss when fine-tuning on BioBERT. Looking at the graph, we can understand that with the increase in the number of epochs, the training loss is constantly decreasing and approaching 0. The validation loss decreases till 8 epochs and then starts to increase. Based on this, we fine-tuned BioBERT for 8 epochs.

We have also tried running LLAMA2 for this transfer learning task. The LLAMA2 model is fine-tuned on the organizational data. The dataset is prepared in such a way that it will have instruction, input, and output, as given in Example 5.2.3. In the test data, the output part will be empty. In the test data, only the sentences with the cause/effect relations are selected. There are a total of 175 sentences in the test data. For each sentence, LLAMA2 will predict the cause and effect phrases. The predicted phrases are then post-processed to extract the sentence and its cause/effect phrase. This is then merged with the gold annotated data to compute the phrase level score. We computed the Jaccard distance, and the cosine similarity between the gold annotated and the predicted phrases. For the zero-shot transfer learning with 3 epochs, we got a Jaccard similarity of 0.50 which means there is a 50% similarity between the gold and the predictions. We got a cosine similarity value of 0.56. The cosine similarity is obtained by computing the vectors of both the gold and the pre-

Table 6.5: Phrase level comparison results of LLAMA2. The model is fine-tuned on Organizational data and tested on medical guidelines data. Jaccard similarity and cosine similarity indicates the average similarity between the gold and the predictions. The F1 score is the comparison between the gold labels and predicted labels.

	Jaccard similarity	Cosine similarity	F1-score
LLAMA2 (3epochs)	0.50	0.56	0.49
LLAMA2 (5epochs)	0.48	0.21	0.64
LLAMA2 (10epochs)	0.48	0.20	0.64

dictions using the Universal Sentence Encoder[150]. The computed values are used to compute the pairwise cosine similarity between two vectors using Scikit-learn [4].

Looking into the results of LLAMA2, we can understand that the predictions are better half of the time. One more observation from the results is that the LLAMA2 predicts one cause and one outcome for all the sentences. This is because the training data (Organizational data) is annotated with one cause and outcome for each sentence. So if there is more than one cause or outcome in the same sentence, the model is not predicting it.

For the sake of the comparison with the results of BERT, DistilBERT, and BioBERT, we converted the phrase-level results into tokens and computed the F1 score. We got an average F1 score of 0.49. Detailed results of the zero-shot transfer learning using LLAMA2 are available in Appendix C Table C.3. From the results, we can understand that LLAMA2 is performing better than the baseline models, which are standard classifiers, but the results of BioBERT are higher compared to LLAMA2.

We tried increasing the number of the epochs to 5 epochs and 10 epochs. With the increase in the number of epochs, the Jaccard similarity and the cosine similarity decreases. However, the F1 score, which compares the gold label with the predicted label, increases with five epochs and remains the same with ten epochs.

[4] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html`

6.2.2 Cross domain few-shot transfer learning

The link between the similarity of training and testing data, and the efficiency of transfer learning was detailed in Section 6.1. The results indicate that a certain degree of resemblance between the training and testing data can enhance the performance of the model. To validate this observation between two vastly diverse dataset domains (organizational and medical), we're proposing a cross-domain few-shot transfer learning. At first, we arbitrarily incorporated 10% of the medical data into the training data (organizational data) and executed all three transformer-based models (BERT, DistilBERT, and BioBERT). We ran all three models to a number of epochs, similar to zero-shot learning. We observed an approximate increase of 10 to 12% in the F1-score for all models. Furthermore, BioBERT displayed superior performance, providing a higher average F1 score of 0.71 within this few-shot learning framework. To evaluate the precision of each model, we implemented a five-fold cross-validation. The outcome revealed an increase in the average F1 score by around 20% compared to the zero-shot learning. Hence, employing few-shot learning with cross-validation resulted in an improved F1 score with all three transformer-based models. We fine-tuned the model for three epochs while implementing five-fold cross-validation. The summary of the results of the few-shot transfer learning is given in Table 6.6. The performance of all three models surpassed that of zero-shot transfer learning. When few-shot transfer learning with cross-validation is implemented, the three transformer-based models exhibit almost equivalent performance. Among these, BioBERT and BERT-based-uncased demonstrated a marginally better performance compared to BERT.

Table 6.6: Summary of the results of cross-domain few-shot transfer learning. Here a five-fold cross-validation was performed, and the smaller part of the split was appended with the training data(organizational data). From the results, we can infer that BioBERT and BERT are performing better than all the other models. Also, compared to zero-shot transfer learning performance, all the models perform better with few-shot transfer learning.

	Precision	Recall	F1-Score	Accuracy
Random forest	0.41	0.41	0.40	0.41
Logistic regression	0.44	0.43	0.38	0.38
DistilBERT	0.78	0.77	0.77	0.80
BERT	0.81	0.80	0.79	0.81
BioBERT	0.81	0.78	0.79	0.81

This chapter provided the findings of our research on identifying the correlation between the training and testing data's similarity and the effectiveness of transfer learning. We corroborated these findings through the application of cross-domain zero-shot and few-shot transfer learning. The results indicate that cross-domain few-shot transfer learning can be a possible solution to address the limited availability of annotated data.

CHAPTER 7: CONCLUSIONS

Throughout this thesis, we tackled the challenges arising from an abundance of Clinical Practice Guidelines in the healthcare sector. We utilized Natural Language Processing techniques to extract causal relations from these medical guidelines. Furthermore, we pioneered a method of inter-sentence level causal sentence classification applied to a corpus of medical guidelines, evaluating the magnitude of the impact from cause to effect. Finally, we showcased the effective use of cross-domain transfer learning to overcome the shortage of annotations.

The main contributions of this dissertation, along with the future directions, are summarized below:

- We developed an automated technique for extracting causalities from annotated corpora of medical guidelines. Additionally, we exhibited the practicality of employing new Large Language Models for causality extraction tasks. With BioBERT, we got an average F1-score of 0.61, whereas with LLAMA2, an average Jaccard distance of 0.40 was obtained.

  We demonstrated the potential for extracting causalities from medical guidelines using a small annotated corpus. The next logical step could involve expanding the corpus through the annotation of more data and creating a benchmark dataset for causality extraction from medical guidelines.

- An inter-sentence level causal sentence classification that can be used to classify the sentences with causal relations across consecutive sentences. We employed a cross-domain inter-sentence level classification between organizational and medical data. An average F1-score of 0.51 was obtained. This work establishes

a baseline for the cross-domain inter-sentence level causal sentence classification.

The accuracy of inter-sentence level classification for medical data could be enhanced with more data. A few-shot approach, similar to what we previously employed in 6.2.2, can be a viable method for achieving this.

- A successful cross-domain intra-sentence level transfer learning between two different domains (organizational and medical). With a few-shot cross-domain transfer learning compared to a zero-shot transfer learning, we got approximately 20% increase in the F1-score.

  – In preparation for the transfer learning, we introduced a novel text-to-knowledge graph pipeline on an annotated corpus of the financial text. For the causality extraction, we got an average F1-score of 0.81, and the results were published in the Information Journal (Published in Information Journal [11] and patent application [35]).

  We employed a basic intra-sentence level explicit causality extraction. This approach could be expanded to incorporate an inter-sentence level and the implementation of implicit causality. Currently, this pipeline is designed specifically for financial text, but it holds potential for extension across multiple domains.

  We showed successful transfer learning with BERT. This can be extended to various LLMs. We can also check the stability of this result by running it for more number of times.

The potential of this research opens up novel dimensions for the health domain, as causality extraction from medical guidelines can enhance clinical decision-making and patient care. This dissertation explored both machine learning and natural language processing techniques for causality extraction. Despite the abundance of causal sentences within these guidelines, automatic extraction is an unexplored field of research.

Also, machine learning models often fail in clinical applications [196] due to the gap between data (both training and testing). In order to avoid this gap, more realistic tests need to be done so that they can be employed for real-world data.

Moreover, the method of quantifying the degree of influence showed promise, as it provided a means of understanding the impact and magnitude of cause to effect. However, future work is still needed to refine these methods and validate their effectiveness, as there is no annotated corpus except the one we introduced in this dissertation.

Furthermore, we explored a cross-domain inter-sentence causal sentence classification in order to amplify our extraction outcomes. Inter-sentence level causality extraction remains a relatively less explored area in research, but utilizing cross-domain inter-sentence causal sentence classification could help expand the inter-sentence level corpus.

This thesis introduced techniques for the automated analysis of causal statements in Clinical Practice Guidelines, enabling an automatic comparison of such discrepancies within these guidelines. This work can be extended in several directions enabling other meta-analyses. For example, further semantic analyses of the content of the guidelines; temporal analysis of recommendations; the examination of geographical variations (e.g. by country) or of contributions of different research centers. Such analyses can potentially improve the methodology of writing clinical practice guidelines.

REFERENCES

[1] H. M. Hapke, H. Lane, and C. Howard, "Natural Language Processing in Action," 2018.

[2] H. Lane, C. Howard, and H. Hapke, *Natural Language Processing in Action*. Manning Publications, 2018.

[3] Natural Language Processing, "Natural Language Processing — Wikipedia, the free encyclopedia," 2020.

[4] D. Jurafsky and J. H. Martin, *Speech & Language Processing3*. retrieved on December, 2022.

[5] H. Hematialam, L. Garbayo, S. Gopalakrishnan, and W. Zadrozny, "Computing Conceptual Distances between Breast Cancer Screening Guidelines: An Implementation of a Near-Peer Epistemic Model of Medical Disagreement," *arXiv preprint arXiv:2007.00709*, 2020.

[6] H. Hematialam, L. Garbayo, S. Gopalakrishnan, and W. W. Zadrozny, "A method for Computing Conceptual Distances between Medical Recommendations: Experiments in Modeling Medical Disagreement," *Applied Sciences*, vol. 11, no. 5, p. 2045, 2021.

[7] O. Hassanzadeh, D. Bhattacharjya, M. Feblowitz, K. Srinivas, M. Perrone, S. Sohrabi, and M. Katz, "Answering Binary Causal Questions Through Large-Scale Text Mining: An Evaluation Using Cause-Effect Pairs from Human Experts.," in *IJCAI*, pp. 5003–5009, 2019.

[8] H. A. Dang, "A study on extracting cause-effect relations and these application for why-question answering," 2021.

[9] S. Gopalakrishnan, S. Padithala, H. Demirhan, and W. Zadrozny, "MDS_UNCC Question Answering System for Biomedical Data with Preliminary Error Analysis.," in *CLEF (Working Notes)*, pp. 231–240, 2021.

[10] S. Gopalakrishnan, V. Chen, G. Hahn-Powell, and B. Tirunagar, "Computer-assisted construct classification of organizational performance concerning different stakeholder groups," *arXiv preprint arXiv:2107.05133*, 2021.

[11] S. Gopalakrishnan, V. Z. Chen, W. Dou, G. Hahn-Powell, S. Nedunuri, and W. Zadrozny, "Text to causal knowledge graph: A framework to synthesize knowledge from unstructured business texts into causal graphs," *Information*, vol. 14, no. 7, p. 367, 2023.

[12] 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.

[13] OpenAI, "GPT-4 technical report," 2023.

[14] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[15] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[16] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.

[17] 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.

[18] J. R. Troche, S. T. Mayne, N. D. Freedman, F. M. Shebl, and C. C. Abnet, "The association between alcohol consumption and lung carcinoma by histological subtype," *American Journal of Epidemiology*, vol. 183, no. 2, pp. 110–121, 2016.

[19] C. Wang, T. Yang, X.-f. Guo, and D. Li, "The associations of fruit and vegetable intake with lung cancer risk in participants with different smoking status: a meta-analysis of prospective cohort studies," *Nutrients*, vol. 11, no. 8, p. 1791, 2019.

[20] R. Girju, "Automatic detection of causal relations for question answering," in *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, pp. 76–83, 2003.

[21] A. Sobrino, C. Puente, and J. A. Olivas, "Extracting answers from causal mechanisms in a medical document," *Neurocomputing*, vol. 135, pp. 53–60, 2014.

[22] C. S. Khoo, S. H. Myaeng, and R. N. Oddy, "Using cause-effect relations in text to improve information retrieval precision," *Information Processing & Management*, vol. 37, no. 1, pp. 119–145, 2001.

[23] Y. Ding, J. Tang, and F. Guo, "Identification of drug-side effect association via multiple information integration with centered kernel alignment," *Neurocomputing*, vol. 325, pp. 211–224, 2019.

[24] S. Zhao, M. Jiang, M. Liu, B. Qin, and T. Liu, "Causaltriad: toward pseudo causal relation discovery and hypotheses generation from medical text data," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 184–193, 2018.

[25] C. S. Khoo, S. Chan, and Y. Niu, "Extracting causal knowledge from a medical database using graphical patterns," in *Proceedings of the 38th annual meeting of the Association for Computational Linguistics*, pp. 336–343, 2000.

[26] G. Dhiman, S. Juneja, W. Viriyasitavat, H. Mohafez, M. Hadizadeh, M. A. Islam, I. El Bayoumy, and K. Gulati, "A novel machine-learning-based hybrid cnn model for tumor identification in medical image processing," *Sustainability*, vol. 14, no. 3, p. 1447, 2022.

[27] Z. Li, F. Liu, L. Antieau, Y. Cao, and H. Yu, "Lancet: a high precision medication event extraction system for clinical text," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 563–567, 2010.

[28] J. Zhang, M. Liu, and Y. Zhang, "Topic-informed neural approach for biomedical event extraction," *Artificial intelligence in medicine*, vol. 103, p. 101783, 2020.

[29] S. Narayanan, K. Mannam, P. Achan, M. V. Ramesh, P. V. Rangan, and S. P. Rajan, "A contextual multi-task neural approach to medication and adverse events identification from clinical text," *Journal of biomedical informatics*, vol. 125, p. 103960, 2022.

[30] X. Dai, S. Karimi, and C. Paris, "Medication and adverse event extraction from noisy text," in *Proceedings of the Australasian Language Technology Association Workshop 2017*, pp. 79–87, 2017.

[31] Q. Wei, Z. Ji, Z. Li, J. Du, J. Wang, J. Xu, Y. Xiang, F. Tiryaki, S. Wu, Y. Zhang, *et al.*, "A study of deep learning approaches for medication and adverse drug event extraction from clinical text," *Journal of the American Medical Informatics Association*, vol. 27, no. 1, pp. 13–21, 2020.

[32] A. Akkasi and M.-F. Moens, "Causal relationship extraction from biomedical text using deep neural models: A comprehensive survey," *Journal of Biomedical Informatics*, vol. 119, p. 103820, 2021.

[33] 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.

[34] I. Reklos and A. Meroño-Peñuela, "Medicause: Causal relation modelling and extraction from medical publications," in *Proceedings of the 1st International Workshop on Knowledge Graph Generation From Text co-located with 19th Extended Semantic Conference (ESWC 2022), Hersonissos, Greece*, vol. 3184, pp. 1–18, 2022.

[35] W. Zadrozny, V. Chen, W. Dou, and S. Gopalakrishnan, "A method and system for integrative causal modeling and transfer," 2022. US Provisional Patent App. 63/429,586.

[36] W. J. Green and M. M. Cheng, "Materiality judgments in an integrated reporting setting: The effect of strategic relevance and strategy map," *Accounting, Organizations and Society*, vol. 73, pp. 1–14, 2019.

[37] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *et al.*, "Knowledge discovery and data mining: Towards a unifying framework.," in *KDD*, vol. 96, pp. 82–88, 1996.

[38] R. Feldman and I. Dagan, "Knowledge discovery in textual databases (KDT).," in *KDD*, vol. 95, pp. 112–117, 1995.

[39] Text Mining, "Text mining — Wikipedia, the free encyclopedia," 2020.

[40] H. R. Bernard and G. Ryan, "Text analysis," *Handbook of methods in cultural anthropology*, vol. 613, 1998.

[41] S. Pawar, G. K. Palshikar, and P. Bhattacharyya, "Relation extraction: A survey," *arXiv preprint arXiv:1712.05191*, 2017.

[42] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[43] 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.

[44] N. Bach and S. Badaskar, "A review of relation extraction," *Literature review for Language and Statistics II*, vol. 2, pp. 1–15, 2007.

[45] B. Waltl, G. Bonczek, and F. Matthes, "Rule-based information extraction: Advantages, limitations, and perspectives," *Jusletter IT (02 2018)*, vol. 4, 2018.

[46] A. Mykowiecka, M. Marciniak, and A. Kupść, "Rule-based information extraction from patients' clinical data," *Journal of Biomedical Informatics*, vol. 42, no. 5, pp. 923–936, 2009.

[47] Y. Zhang, H. Fei, and P. Li, "End-to-end distantly supervised information extraction with retrieval augmentation," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2449–2455, 2022.

[48] G. Jacobs and V. Hoste, "Sentivent: enabling supervised information extraction of company-specific events in economic and financial news," *Language Resources and Evaluation*, vol. 56, no. 1, pp. 225–257, 2022.

[49] B. B. Dalvi, W. W. Cohen, and J. Callan, "Websets: Extracting sets of entities from the web using unsupervised information extraction," in *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 243–252, 2012.

[50] M. Michelson and C. A. Knoblock, "Unsupervised information extraction from unstructured, ungrammatical data sources on the world wide web," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 10, pp. 211–226, 2007.

[51] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka Jr, and T. M. Mitchell, "Coupled semi-supervised learning for information extraction," in *Proceedings of the third ACM international conference on Web search and data mining*, pp. 101–110, 2010.

[52] Y. Chong, Y. Ding, Q. Yan, and S. Pan, "Graph-based semi-supervised learning: A review," *Neurocomputing*, vol. 408, pp. 216–230, 2020.

[53] R. Zhang and N. El-Gohary, "A deep neural network-based method for deep information extraction using transfer learning strategies to support automated compliance checking," *Automation in Construction*, vol. 132, p. 103834, 2021.

[54] N. Zhang, H. Ye, S. Deng, C. Tan, M. Chen, S. Huang, F. Huang, and H. Chen, "Contrastive information extraction with generative transformer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3077–3088, 2021.

[55] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey of web information extraction systems," *IEEE transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1411–1428, 2006.

[56] 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.

[57] D. Garcia *et al.*, "Coatis, an nlp system to locate expressions of actions connected by causality links," in *International Conference on Knowledge Engineering and Knowledge Management*, pp. 347–352, Springer, 1997.

[58] C. Khoo, S. Chan, and Y. Niu, "The many facets of the cause-effect relation," in *The Semantics of Relationships*, pp. 51–70, Springer, 2002.

[59] K. Radinsky, S. Davidovich, and S. Markovitch, "Learning causality for news events prediction," in *Proceedings of the 21st International Conference on World Wide Web*, pp. 909–918, 2012.

[60] A. Ittoo and G. Bouma, "Minimally-supervised learning of domain-specific causal relations using an open-domain corpus as knowledge base," *Data & Knowledge Engineering*, vol. 88, pp. 142–163, 2013.

[61] N. Kang, B. Singh, C. Bui, Z. Afzal, E. M. van Mulligen, and J. A. Kors, "Knowledge-based extraction of adverse drug events from biomedical text," *BMC Bioinformatics*, vol. 15, no. 1, pp. 1–8, 2014.

[62] C. S. Khoo, J. Kornfilt, R. N. Oddy, and S. H. Myaeng, "Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing," *Literary and Linguistic Computing*, vol. 13, no. 4, pp. 177–186, 1998.

[63] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." 2017.

[64] 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*, pp. 55–60, 2014.

[65] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," *arXiv preprint arXiv:2003.07082*, 2020.

[66] I. Keskes, F. B. Zitoune, and L. H. Belguith, "Learning explicit and implicit arabic discourse relations," *Journal of King Saud University-Computer and Information Sciences*, vol. 26, no. 4, pp. 398–416, 2014.

[67] C. Pechsiri, A. Kawtrakul, and R. Piriyakul, "Mining causality knowledge from textual data.," in *Artificial Intelligence and Applications*, pp. 85–90, 2006.

[68] J.-H. Oh, K. Torisawa, C. Hashimoto, M. Sano, S. De Saeger, and K. Ohtake, "Why-question answering using intra-and inter-sentential causal relations," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1733–1743, 2013.

[69] Y. Xu, L. Mou, G. Li, Y. Chen, H. Peng, and Z. Jin, "Classifying relations via long short term memory networks along shortest dependency paths," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1785–1794, 2015.

[70] Z. Li, Q. Li, X. Zou, and J. Ren, "Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings," *Neurocomputing*, vol. 423, pp. 207–219, 2021.

[71] L. Wang, Z. Cao, G. De Melo, and Z. Liu, "Relation classification via multi-level attention cnns," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1298–1307, 2016.

[72] M. Kyriakakis, I. Androutsopoulos, A. Saudabayev, *et al.*, "Transfer learning for causal sentence detection," *arXiv preprint arXiv:1906.07544*, 2019.

[73] F. Li, M. Zhang, G. Fu, and D. Ji, "A neural joint model for entity and relation extraction from biomedical text," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–11, 2017.

[74] H. Man, M. Nguyen, and T. Nguyen, "Event causality identification via generation of important context words," in *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pp. 323–330, 2022.

[75] A. Balashankar, S. Chakraborty, S. Fraiberger, and L. Subramanian, "Identifying predictive causal factors from news streams," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2338–2348, 2019.

[76] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "Flair: An easy-to-use framework for state-of-the-art NLP," in *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics (demonstrations)*, pp. 54–59, 2019.

[77] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[78] J. Fischbach, T. Springer, J. Frattini, H. Femmer, A. Vogelsang, and D. Mendez, "Fine-grained causality extraction from natural language requirements using recursive neural tensor networks," in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pp. 60–69, IEEE, 2021.

[79] S. J. Sheikh, S. Haider, and A. H. Levis, "On semi-automated extraction of causal networks from raw text," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106189, 2023.

[80] M. Wu, Q. Zhang, C. Wu, and G. Wang, "End-to-end multi-granulation causality extraction model," *Digital Communications and Networks*, 2023.

[81] T. Schick and H. Schütze, "Exploiting cloze questions for few shot text classification and natural language inference," *arXiv preprint arXiv:2001.07676*, 2020.

[82] N. Zhang, L. Li, X. Chen, S. Deng, Z. Bi, C. Tan, F. Huang, and H. Chen, "Differentiable prompt makes pre-trained language models better few-shot learners," *arXiv preprint arXiv:2108.13161*, 2021.

[83] X. Chen, N. Zhang, L. Li, X. Xie, S. Deng, C. Tan, F. Huang, L. Si, and H. Chen, "Lightner: A lightweight generative framework with prompt-guided attention for low-resource NER," *arXiv preprint arXiv:2109.00720*, 2021.

[84] L. Cui, Y. Wu, J. Liu, S. Yang, and Y. Zhang, "Template-based named entity recognition using BART," *arXiv preprint arXiv:2106.01760*, 2021.

[85] J. Liu, Z. Zhang, Z. Guo, L. Jin, X. Li, K. Wei, and X. Sun, "Kept: Knowledge enhanced prompt tuning for event causality identification," *Knowledge-Based Systems*, vol. 259, p. 110064, 2023.

[86] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz, "SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, (Uppsala, Sweden), pp. 33–38, Association for Computational Linguistics, July 2010.

[87] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, "The Penn Discourse Treebank 2.0.," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

[88] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning, "Position-aware attention and supervised data improve slot filling," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pp. 35–45, 2017.

[89] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, and M. Sun, "DocRED: A large-scale document-level relation extraction dataset," in *Proceedings of ACL 2019*, 2019.

[90] Z. Song and S. Strassel, "Entity translation and alignment in the ACE-07 ET task," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

[91] A. Mandya, D. Bollegala, F. Coenen, K. Atkinson, and T. Declerck, "A dataset for inter-sentence relation extraction using distant supervision," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 1559–1565, 2018.

[92] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003–1011, 2009.

[93] F. A. Tan, A. Hürriyetoğlu, T. Caselli, N. Oostdijk, T. Nomoto, H. Hettiarachchi, I. Ameer, O. Uca, F. F. Liza, and T. Hu, "The causal news corpus: Annotating causal relations in event sentences from news," *arXiv preprint arXiv:2204.11714*, 2022.

[94] C. Walker, S. Strassel, J. Medero, and K. Maeda, "ACE 2005 multilingual training corpus," *Linguistic Data Consortium, Philadelphia*, vol. 57, p. 45, 2006.

[95] D. Mariko, H. Abi Akl, K. Trottier, and M. El-Haj, "The financial causality extraction shared task (FinCausal 2022)," in *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pp. 105–107, 2022.

[96] D. Buscaldi, A.-K. Schumann, B. Qasemizadeh, H. Zargayouna, and T. Charnois, "Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers," in *Proceedings of the 12th International Workshop on Semantic Evaluation*, pp. 679–688, 2018.

[97] M. J. Field, K. N. Lohr, *et al.*, "Clinical practice guidelines," *Directions for a new program*, p. 1990, 1990.

[98] E. Steinberg, S. Greenfield, D. M. Wolman, M. Mancher, R. Graham, *et al.*, *Clinical practice guidelines we can trust*. National Academies Press, 2011.

[99] B. T. S. S. I. G. Network *et al.*, "British guideline on the management of asthma," *Thorax*, vol. 69, no. Suppl 1, pp. i1–i192, 2008.

[100] U. P. S. T. Force *et al.*, "Screening for breast cancer: Recommendation statement," *Rockville, MD: Agency for Healthcare Research and Quality (AHRQ)*, 2011.

[101] A. Wasylewicz and A. Scheepers-Hoeks, "Clinical decision support systems," *Fundamentals of clinical data science*, pp. 153–169, 2019.

[102] A. R. Ertle, E. Campbell, and W. R. Hersh, "Automated application of clinical practice guidelines for asthma management.," in *Proceedings of the AMIA Annual Fall Symposium*, p. 552, American Medical Informatics Association, 1996.

[103] M. Taboada, M. Meizoso, D. Martínez, D. Riano, and A. Alonso, "Combining open-source natural language processing tools to parse clinical practice guidelines," *Expert Systems*, vol. 30, no. 1, pp. 3–11, 2013.

[104] K. Kaiser and S. Miksch, "Supporting the abstraction of clinical practice guidelines using information extraction," in *International Conference on Application of Natural Language to Information Systems*, pp. 304–311, Springer, 2010.

[105] W. Chunhua, P. R. Payne, M. Velez, S. B. Johnson, and S. Bakken, "Towards symbiosis in knowledge representation and natural language processing for structuring clinical practice guidelines," *Studies in health technology and informatics*, vol. 201, p. 461, 2014.

[106] L. B. Fazlic, A. Hallawa, A. Schmeink, A. Peine, L. Martin, and G. Dartmann, "A novel NLP-fuzzy system prototype for information extraction from medical guidelines," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1025–1030, IEEE, 2019.

[107] S. S. Graham, Z. P. Majdik, J. B. Barbour, and J. F. Rousseau, "Associations Between Aggregate NLP-extracted Conflicts of Interest and Adverse Events By Drug Product," *Studies in health technology and informatics*, vol. 290, p. 405, 2022.

[108] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh, "Calibrate before use: Improving few-shot performance of language models," in *International Conference on Machine Learning*, pp. 12697–12706, PMLR, 2021.

[109] M. Y. Landolsi, L. Hlaoua, and L. Ben Romdhane, "Information extraction from electronic medical documents: state of the art and future research directions," *Knowledge and Information Systems*, vol. 65, no. 2, pp. 463–516, 2023.

[110] H. Hematialam and W. Zadrozny, "Identifying condition-action statements in medical guidelines using domain-independent features," *arXiv preprint arXiv:1706.04206*, 2017.

[111] 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*, pp. 26–38, Springer, 2013.

[112] H. Hematialam, *Knowledge Extraction and Analysis of Medical Text with Particular Emphasis on Medical Guidelines*. PhD thesis, The University of North Carolina at Charlotte, 2021.

[113] M. Hussain, J. Hussain, M. Sadiq, A. U. Hassan, and S. 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.

[114] Q. Xie, J. A. Bishop, P. Tiwari, and S. Ananiadou, "Pre-trained language models with domain knowledge for Biomedical extractive summarization," *Knowledge-Based Systems*, p. 109460, 2022.

[115] D. R. Schlegel, K. Gordon, C. Gaudioso, and M. Peleg, "Clinical tractor: A framework for automatic natural language understanding of clinical practice guidelines," in *AMIA Annual Symposium Proceedings*, vol. 2019, p. 784, American Medical Informatics Association, 2019.

[116] 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*, vol. 109, p. 103526, 2020.

[117] R. Tang, X. Han, X. Jiang, and X. Hu, "Does synthetic data generation of llms help clinical text mining?," *arXiv preprint arXiv:2303.04360*, 2023.

[118] S. Henry, K. Buchan, M. Filannino, A. Stubbs, and O. Uzuner, "2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records," *Journal of the American Medical Informatics Association*, vol. 27, no. 1, pp. 3–12, 2020.

[119] A. P. Kurniati, E. Rojas, D. Hogg, G. Hall, and O. A. Johnson, "The assessment of data quality issues for process mining in healthcare using medical information mart for intensive care iii, a freely available e-health record database," *Health Informatics Journal*, vol. 25, no. 4, pp. 1878–1893, 2019.

[120] M. Filannino, A. Stubbs, and Ö. Uzuner, "Symptom severity prediction from neuropsychiatric clinical records: Overview of 2016 cegs n-grid shared tasks track 2," *Journal of Biomedical Informatics*, vol. 75, pp. S62–S70, 2017.

[121] A. Jagannatha, F. Liu, W. Liu, and H. Yu, "Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (made 1.0)," *Drug safety*, vol. 42, no. 1, pp. 99–111, 2019.

[122] A. Miranda-Escalada, E. Farré, and M. Krallinger, "Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results," in *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*, 2020.

[123] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, "Biocreative v cdr task corpus: a resource for chemical disease relation extraction," *Database*, vol. 2016, 2016.

[124] C. Mihăilă, T. Ohta, S. Pyysalo, and S. Ananiadou, "Biocause: Annotating and analysing causality in the biomedical domain," *BMC bioinformatics*, vol. 14, pp. 1–18, 2013.

[125] R. Prasad, S. McRoy, N. Frid, A. Joshi, and H. Yu, "The Biomedical Discourse Relation Bank," *BMC Bioinformatics*, vol. 12, no. 1, pp. 1–18, 2011.

[126] S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski, "Bioinfer: a corpus for information extraction in the biomedical domain," *BMC bioinformatics*, vol. 8, pp. 1–24, 2007.

[127] J.-D. Kim, T. Ohta, and J. Tsujii, "Corpus annotation for mining biomedical events from literature," *BMC bioinformatics*, vol. 9, pp. 1–25, 2008.

[128] J. D. Gallow, "The Metaphysics of Causation," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta and U. Nodelman, eds.), Metaphysics Research Lab, Stanford University, Fall 2022 ed., 2022.

[129] M. A. Kabir, A. Almulhim, X. Luo, and M. Al Hasan, "Informative causality extraction from medical literature via dependency-tree–based patterns," *Journal of Healthcare Informatics Research*, pp. 1–22, 2022.

[130] M. Fajcik, M. Singh, J. Zuluaga-Gomez, E. Villatoro-Tello, S. Burdisso, P. Motlicek, and P. Smrz, "Idiapers@ causal news corpus 2022: Extracting cause-effect-signal triplets via pre-trained autoregressive language model," *arXiv preprint arXiv:2209.03891*, 2022.

[131] K. W. Davidson, M. J. Barry, C. M. Mangione, M. Cabana, A. B. Caughey, E. M. Davis, K. E. Donahue, C. A. Doubeni, M. Kubik, L. Li, *et al.*, "Screening

for gestational diabetes: US Preventive Services Task Force recommendation statement," *JAMA*, vol. 326, no. 6, pp. 531–538, 2021.

[132] J. Mills and S. Mohnot, "Screening for gestational diabetes," *American Family Physician*, vol. 104, no. 6, pp. 641–642, 2021.

[133] W. C. Members, J. S. Lawton, J. E. Tamis-Holland, S. Bangalore, E. R. Bates, T. M. Beckie, J. M. Bischoff, J. A. Bittl, M. G. Cohen, J. M. DiMaio, *et al.*, "2021 acc/aha/scai guideline for coronary artery revascularization: A report of the american college of cardiology/american heart association joint committee on clinical practice guidelines," *Journal of the American College of Cardiology*, vol. 79, no. 2, pp. e21–e129, 2022.

[134] D. L. Monticciolo, M. S. Newell, L. Moy, B. Niell, B. Monsees, and E. A. Sickles, "Breast cancer screening in women at higher-than-average risk: recommendations from the acr," *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 408–414, 2018.

[135] J. Fischbach, J. Frattini, A. Spaans, M. Kummeth, A. Vogelsang, D. Mendez, and M. Unterkalmsteiner, "Automatic detection of causality in requirement artifacts: the cira approach," in *Requirements Engineering: Foundation for Software Quality: 27th International Working Conference, REFSQ 2021, Essen, Germany, April 12–15, 2021, Proceedings 27*, pp. 19–36, Springer, 2021.

[136] "2. classification and diagnosis of diabetes: Standards of medical care in diabetes—2020, American Diabetes Association," *Diabetes care*, vol. 43, no. Supplement_1, pp. S14–S31, 2020.

[137] "ACOG practice bulletin, mellitus, gestational diabetes," *ACOG: Washington, DC, USA*, 2018.

[138] I. Blumer, E. Hadar, D. R. Hadden, L. Jovanovič, J. H. Mestman, M. H. Murad, and Y. Yogev, "Diabetes and pregnancy: An Endocrine society clinical practice guideline," *The journal of clinical endocrinology & Metabolism*, vol. 98, no. 11, pp. 4227–4249, 2013.

[139] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions*. john wiley & sons, 2013.

[140] G. Hripcsak and A. S. Rothschild, "Agreement, the f-measure, and reliability in information retrieval," *Journal of the American medical informatics association*, vol. 12, no. 3, pp. 296–298, 2005.

[141] P. Thompson, S. A. Iqbal, J. McNaught, and S. Ananiadou, "Construction of an annotated corpus to support biomedical information extraction," *BMC bioinformatics*, vol. 10, pp. 1–19, 2009.

[142] F. P. Miller, A. F. Vandome, and J. McBrewster, "Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau? levenshtein distance, spell checker, hamming distance," 2009.

[143] R. Real and J. M. Vargas, "The probabilistic basis of jaccard's index of similarity," *Systematic biology*, vol. 45, no. 3, pp. 380–385, 1996.

[144] A. Barbaresi, "Trafilatura: A web scraping library and command-line tool for text discovery and extraction," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 122–131, 2021.

[145] S. Bird, "Nltk: the natural language toolkit," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pp. 69–72, 2006.

[146] E. Barrett, J. Paradis, and L. C. Perelman, "The mayfield handbook of technical & scientific writing," *Mountain View, CA: Mayfield Company*, 1998.

[147] P. Shi and J. Lin, "Simple bert models for relation extraction and semantic role labeling," *arXiv preprint arXiv:1904.05255*, 2019.

[148] C. Lin, T. Miller, D. Dligach, S. Bethard, and G. Savova, "A bert-based universal model for both within-and cross-sentence clinical temporal relation extraction," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 65–71, 2019.

[149] T. Nayak, S. Sharma, Y. Butala, K. Dasgupta, P. Goyal, and N. Ganguly, "A generative approach for financial causality extraction," in *Companion Proceedings of the Web Conference 2022*, pp. 576–578, 2022.

[150] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[151] C. Chan, J. Cheng, W. Wang, Y. Jiang, T. Fang, X. Liu, and Y. Song, "Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations," *arXiv preprint arXiv:2304.14827*, 2023.

[152] A. Gordon, Z. Kozareva, and M. Roemmele, "Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning," in ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 394–398, 2012.

[153] L. Du, X. Ding, K. Xiong, T. Liu, and B. Qin, "e-care: a new dataset for exploring explainable causal reasoning," *arXiv preprint arXiv:2205.05849*, 2022.

[154] I. Gusev and A. Tikhonov, "Headlinecause: A dataset of news headlines for detecting causalities," *arXiv preprint arXiv:2108.12626*, 2021.

[155] E. Hinkel, "The use of modal verbs as a reflection of cultural values," *TESOL quarterly*, vol. 29, no. 2, pp. 325–343, 1995.

[156] C. Ma, Y. Zhang, and M. Zhang, "Tree Kernel-based Protein-Protein interaction extraction considering both modal verb phrases and appositive dependency features," in *Proceedings of the 24th International Conference on World Wide Web*, pp. 655–660, 2015.

[157] S. B. de Vroe, L. Guillou, M. Stanojević, N. McKenna, and M. Steedman, "Modality and negation in event extraction," *arXiv preprint arXiv:2109.09393*, 2021.

[158] J. Sun, S. Huang, and C. Wei, "A BERT-based deontic logic learner," *Information Processing & Management*, vol. 60, no. 4, p. 103374, 2023.

[159] C.-X. Wan and B. Li, "Financial causal sentence recognition based on BERT-CNN text classification," *The Journal of Supercomputing*, pp. 1–25, 2022.

[160] F. A. Tan, D. Hazarika, S. K. Ng, S. Poria, and R. Zimmermann, "Causal augmentation for causal sentence classification," in *Proceedings of the First Workshop on Causal Inference and NLP*, pp. 1–20, 2021.

[161] F. A. Tan, H. Hettiarachchi, A. Hürriyetoğlu, T. Caselli, O. Uca, F. F. Liza, and N. Oostdijk, "Event Causality Identification with Causal News Corpus–Shared Task 3, CASE 2022," *arXiv preprint arXiv:2211.12154*, 2022.

[162] X. Jin, X. Wang, X. Luo, S. Huang, and S. Gu, "Inter-sentence and implicit causality extraction from chinese corpus," in *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I 24*, pp. 739–751, Springer, 2020.

[163] D.-S. Chang and K.-S. Choi, "Causal relation extraction using cue phrase and lexical pair probabilities," in *International conference on natural language processing*, pp. 61–70, Springer, 2004.

[164] L. Gao, P. K. Choubey, and R. Huang, "Modeling document-level causal structures for event causal relation identification," 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)*, pp. 1808–1817, 2019.

[165] Q. Cao, X. Hao, H. Ren, W. Xu, S. Xu, and C. J. Asiedu, "Graph attention network based detection of causality for textual emotion-cause pair," *World Wide Web*, vol. 26, no. 4, pp. 1731–1745, 2023.

[166] M. T. Phu and T. H. Nguyen, "Graph convolutional networks for event causality identification with rich document-level structures," in *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 3480–3490, 2021.

[167] G. Hahn-Powell, D. Bell, M. A. Valenzuela-Escárcega, and M. Surdeanu, "This before that: Causal precedence in the biomedical domain," *arXiv preprint arXiv:1606.08089*, 2016.

[168] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of statistical planning and inference*, vol. 90, no. 2, pp. 227–244, 2000.

[169] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[170] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[171] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[172] L. Duan, I. W. Tsang, and D. Xu, "Domain transfer multiple kernel learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 465–479, 2012.

[173] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1717–1724, 2014.

[174] T. Tommasi, F. Orabona, and B. Caputo, "Safety in numbers: Learning categories from few examples with multi model knowledge transfer," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3081–3088, IEEE, 2010.

[175] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," in *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 1855–1862, IEEE, 2010.

[176] M. Long, J. Wang, G. Ding, S. J. Pan, and S. Y. Philip, "Adaptation regularization: A general framework for transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1076–1089, 2013.

[177] R. Xia, C. Zong, X. Hu, and E. Cambria, "Feature ensemble plus sample selection: domain adaptation for sentiment classification," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 10–18, 2013.

[178] P. Prettenhofer and B. Stein, "Cross-language text classification using structural correspondence learning," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 1118–1127, 2010.

[179] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 120–128, 2006.

[180] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. Le, "Rethinking pre-training and self-training," *Advances in neural information processing systems*, vol. 33, pp. 3833–3845, 2020.

[181] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[182] H. Hematialam and W. W. Zadrozny, "Identifying Condition-action Statements in Medical Guidelines: Three Studies using Machine Learning and Domain Adaptation," 2021.

[183] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets," *arXiv preprint arXiv:1906.05474*, 2019.

[184] T. Miller, E. Laparra, and S. Bethard, "Domain adaptation in practice: Lessons from a real-world information extraction pipeline," in *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pp. 105–110, 2021.

[185] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[186] Transfer Learning, "Transfer learning — Wikipedia, the free encyclopedia," 2022. [Online].

[187] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International statistical review*, vol. 70, no. 3, pp. 419–435, 2002.

[188] C. D. Manning, *An introduction to information retrieval*. Cambridge university press, 2009.

[189] D. Li, H. Fei, S. Ren, and P. Li, "A deep decomposable model for disentangling syntax and semantics in sentence representation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4300–4310, 2021.

[190] J. H. Martin, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall, 2009.

[191] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger, "Adversarial deep averaging networks for cross-lingual sentiment classification," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 557–570, 2018.

[192] H. Al Kuwatly, M. Wich, and G. Groh, "Identifying and measuring annotator bias based on annotators' demographic characteristics," in *Proceedings of the fourth workshop on online abuse and harms*, pp. 184–190, 2020.

[193] 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.

[194] C. Lyu, T. Ji, Q. Sun, and L. Zhou, "DCU-lorcan at fincausal 2022: Span-based causality extraction from financial documents using pre-trained language models," in *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pp. 116–120, 2022.

[195] J. T. VanSchaik, P. Jain, A. Rajapuri, B. Cheriyan, T. P. Thyvalikakath, and S. Chakraborty, "Using transfer learning-based causality extraction to mine latent factors for sjögren's syndrome from biomedical literature," *Heliyon*, vol. 9, no. 9, 2023.

[196] C. Schmidt, "Md anderson breaks with ibm watson, raising questions about artificial intelligence in oncology," 2017.

[197] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, *et al.*, "Large language models encode clinical knowledge," *arXiv preprint arXiv:2212.13138*, 2022.

[198] C. Si, Z. Gan, Z. Yang, S. Wang, J. Wang, J. Boyd-Graber, and L. Wang, "Prompting gpt-3 to be reliable," *arXiv preprint arXiv:2210.09150*, 2022.

[199] J. Gao, X. Ding, B. Qin, and T. Liu, "Is chatgpt a good causal reasoner? a comprehensive evaluation," *arXiv preprint arXiv:2305.07375*, 2023.

APPENDIX A: Results of causal sentence classification

This chapter summarizes the results of the causal sentence classification. The average results are given in the thesis and the detailed results are given in this appendix.

A.1 Causal sentence classification on Organizational data

Table A.1: Summary of BERT's performance for the causal sentence classification. Here, for the positive sentences, we use the sentences with causal relations from our data. For the negative sample, we have sentences that do not contain causal relations from our data and a random sample of sentences from Twitter data that do not contain causal triggers. We merge the data to form the negative sample.

	Precision	Recall	F1
Class 0: Negative class with all cases of a causal trigger where the sentence does not contain a causal relation and random sample of sentences without causal relations and without causal triggers	0.91	0.86	0.88
Class 1: Positive class, consisting of sentences that contain causal relations	0.86	0.91	0.89

Table A.2: Summary of BERT's performance for the causal sentence classification. Here, for the positive sentences, we use sentences with causal relations from our data. For the negative sample, we have sentences that do not contain causal relations selected from our data. P stands for Precision, R for Recall and F1 for F1-score

	Precision	Recall	F1-score
Class 0: Negative class with all cases of a causal trigger where the sentence does not contain a causal relation	0.91	0.83	0.87
Class 1: Positive class, which consists of sentences that contain causal relations	0.85	0.92	0.88

Table A.3: Summary of the causal sentence classification on medical text using Logistic regression.

	Precision	Recall	F1-score	Support
Non-Causal	0.96	0.73	0.83	292
Causal	0.78	0.97	0.87	292
Accuracy			0.85	584
Macro average	0.87	0.85	0.85	584
Weighted average	0.87	0.85	0.85	584

Table A.4: Summary of the causal sentence classification using DistilBERT.

	Precision	Recall	F1-score	Support
Non-Causal	0.98	0.97	0.97	292
Causal	0.97	0.98	0.97	292
Accuracy			0.97	584
Macro average	0.97	0.97	0.97	584
Weighted average	0.97	0.97	0.97	584

A.2 Causal sentence classification on Medical data

This section summarizes the results of the causal sentence classification on the medical data. The detailed results for the top-performing model are given in the thesis, and the detailed results of all other models are summarized in Table A.3, A.4.

A.3 Causal sentence classification on Causal News Corpus

For the causal sentence classification on medical guidelines text using GPT-4, we used a four-shot prompting. We particularly chose a four-shot prompting based on our similar work on the Causal News Corpus dataset. We noticed the lack of performance improvement after four prompts. The results of various prompt sizes are given in Table A.5.

Table A.5: Results of GPT-4 with prompting vs. fine-tuned BERT for Causal Event Classification task (on the Causal News Corpus). Note the lack of GPT-4 improvement with more prompts.

	Recall	Precision	F1	Accuracy	MCC
GPT-4 zero-shot prompting	0.641	0.613	0.627	0.625	0.25
GPT-4 two-shot prompting	0.716	0.670	0.692	0.687	0.376
GPT-4 four-shot prompting	0.653	0.701	0.677	0.693	0.386
GPT-4 six-shot prompting	0.745	0.671	0.707	0.696	0.395
GPT-4 fourteen-shot prompting	0.711	0.687	0.699	0.699	0.398
BERT-base, fine-tuned	0.861	0.780	0.818	0.812	0.628

APPENDIX B: Results of Causality extraction

B.1 Causality extraction from Organizational data

In recent times, prompting large language models has given state-of-the-art performing results for many NLP tasks [197, 198]. We tried a few-shot prompting of GPT-3 on a sample of 100 sentences from our dataset. The model's results are summarized in Table B.1. At the time of running these experiments, we did not have access to GPT-4.

Table B.1: Few-shot prompting of GPT-3.5 on the organizational causality extraction dataset. This result was on a sample of 100 sentences from the dataset.

	Precision	Recall	F1-Score
Cause	0.49	0.28	0.36
Causal trigger	0.05	0.05	0.05
Effect	0.47	0.38	0.42

From Table B.1 it can be observed that the Large Language Model GPT-3.5 performed poorly on a sample of 100 sentences from our data. [199] discusses the performance of ChatGPT for causal reasoning and causal interpretation. Their experiments showed that ChatGPT was not a good causal reasoner, which our results also indicate.

B.2 Causality extraction from medical data

The graph representing the training loss and the validation loss or DistilBERT and BERT are given below in Figure B.1 and Figure B.2. The results of our best-performing model BioBERT are given in the thesis.

The results of the DistilBERT and the BERT's performance are summarized in Table B.2, B.3.

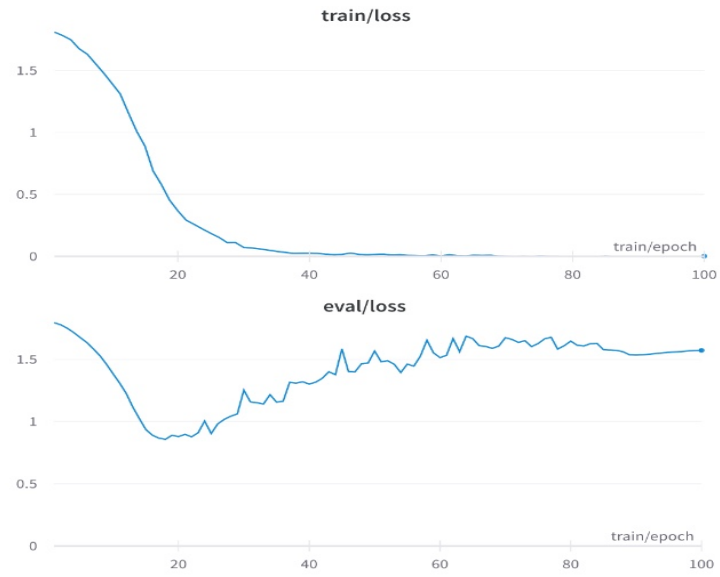


Figure B.1: Graph showing the train and validation loss when fine-tuning on DistilBERT. Looking at the graph, we can understand that with the increase in the number of epochs, the training loss is constantly decreasing and approaching 0. The validation loss decreases till 18 epochs and then starts to increase. Based on this, we fine-tuned DistilBERT for 18 epochs.

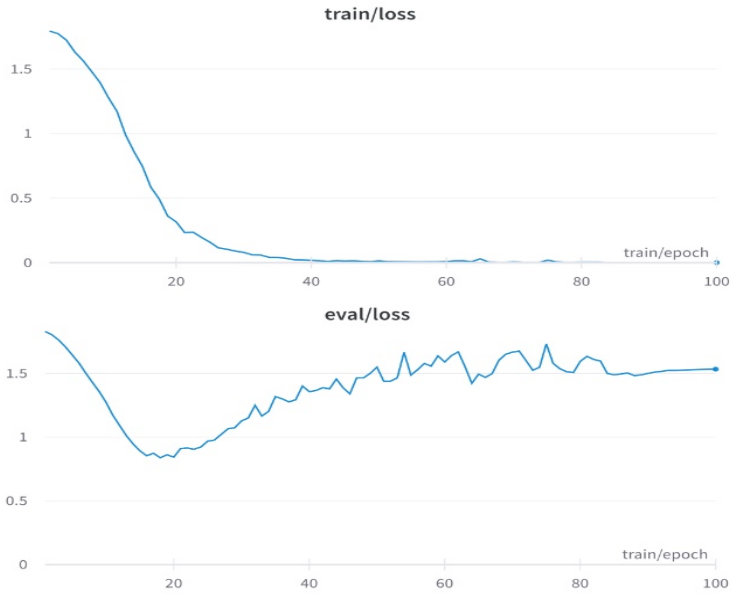


Figure B.2: Graph showing the train and validation loss when fine-tuning on BERT. Looking at the graph, we can understand that with the increase in the number of epochs, the training loss is constantly decreasing and approaching 0. The validation loss decreases till 20 epochs and then starts to increase. Based on this, we fine-tuned BERT for 20 epochs.

Table B.2: Summary of causality extraction results on the medical data using DistilBERT. Each token in the text was assigned a label Signal(S), Effect(E), Other(O), Cause(C), Condition(CO), and Action(A). The results are obtained by splitting the manually annotated data into train and test data. Note: there are very few samples in the entire dataset for the signal, and the F1-score for a signal is 0 with all the models.

	Precision	Recall	F1-score	Support
S	1.00	0.00	0.00	31
E	0.71	0.74	0.72	696
O	0.70	0.73	0.72	1186
C	0.67	0.57	0.62	411
CO	0.73	0.65	0.69	717
A	0.66	0.74	0.69	838
Accuracy			0.69	3879
Macro average	0.74	0.57	0.57	3879
Weighted average	0.70	0.69	0.69	3879

Table B.3: Summary of causality extraction results on the medical data using BERT. Each token in the text was assigned a label Signal(S), Effect(E), Other(O), Cause(C), Condition(CO), and Action(A). The results are obtained by splitting the manually annotated data into train and test data. Note: there are very few samples in the entire dataset for the signal, and the F1-score for a signal is 0 with all the models.

	Precision	Recall	F1-score	Support
S	1.00	0.00	0.00	31
E	0.77	0.76	0.76	696
O	0.75	0.75	0.75	1186
C	0.69	0.59	0.63	411
CO	0.71	0.68	0.70	717
A	0.71	0.83	0.76	838
Accuracy			0.73	3879
Macro average	0.77	0.60	0.60	3879
Weighted average	0.73	0.73	0.73	3879

Table B.4: Summary of causality extraction results on the medical data using GPT-4. Each token in the text was assigned a label Signal(S), Effect(E), Other(O), Cause(C), Condition(CO), and Action(A). Here four-shot prompting is used. Expect the data that are used for prompting the rest of the data was used as test data.

	Precision	Recall	F1-score	Support
S	0.00	0.00	0.00	75
E	0.78	0.36	0.49	1805
O	0.49	0.29	0.37	2499
C	0.24	0.83	0.37	1080
CO	0.46	0.06	0.11	1584
A	0.50	0.69	0.58	2126
Accuracy			0.42	9169
Macro average	0.41	0.37	0.32	9169
Weighted average	0.51	0.42	0.39	9169

B.3 LLAMA2

In this section, we document the command we use for fine-tuning the LLAMA2 model. We fine-tuned the model using HuggingFace autotrain. We also present the detailed results of causality extraction using LLAMA2 in Table B.5. The average F1 score, along with the phrase level distance (Jaccard distance and cosine similarity) score, are reported in Chapter 5.

The autotrain command which we used for fine-tuning LLAMA2:

```
!autotrain llm --train --project\_name 'llama2-causalityextractionGD'  \
--model yahma/llama-7b-hf --data_path /content/data/ --use_peft --\use\_int4  \
--learning\_rate 2e-4 --train\_batch\_size 2 \
--num\_train\_epochs 3 --push_to_hub --repo\_id \
Seetha/llama2-causalityextractionGD \
--trainer sft > training.log &
```

Table B.5: Summary of causality extraction results on the medical data using LLAMA2. Each token in the text was assigned a label Signal(S), Effect(E), Other(O), Cause(C), Condition(CO), and Action(A). The gestational diabetes data is split into train and test. The model is fine-tuned on the training data and tested on the test data. The overall accuracy was 0.43. The baselines are shown in Table 5.5

	Precision	Recall	F1-score
S	0.00	0.00	0.00
E	0.28	0.58	0.38
O	0.37	0.78	0.50
C	0.39	0.54	0.45
CO	0.71	0.25	0.37
A	0.86	0.29	0.44
Macro average	0.43	0.41	0.36
Weighted average	0.61	0.43	0.42

APPENDIX C: Causal transfer learning from organizational data to medical data

This chapter summarizes the detailed results of the transfer learning. The detailed results of the best-performing model are summarized in Chapter 6.

C.1 Zero-shot transfer learning

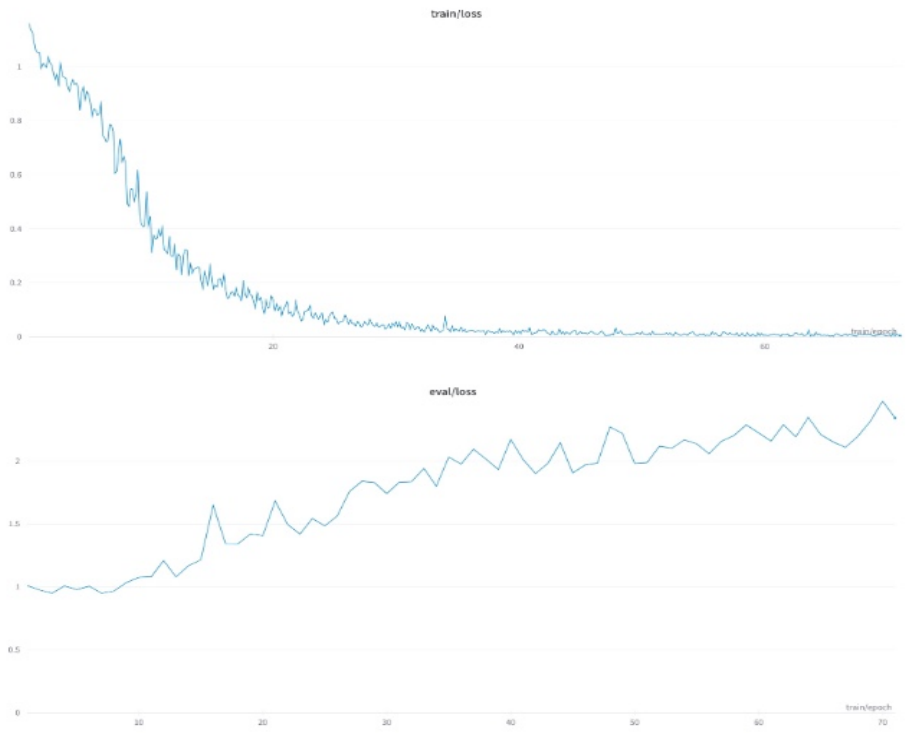


Figure C.1: Graph showing the train and validation loss when fine-tuning on BERT. Looking at the graph, we can understand that with the increase in the number of epochs, the training loss is constantly decreasing and approaching 0. The validation loss decreases till 7 epochs and then starts to increase. Based on this, we fine-tuned DistilBERT for 7 epochs.

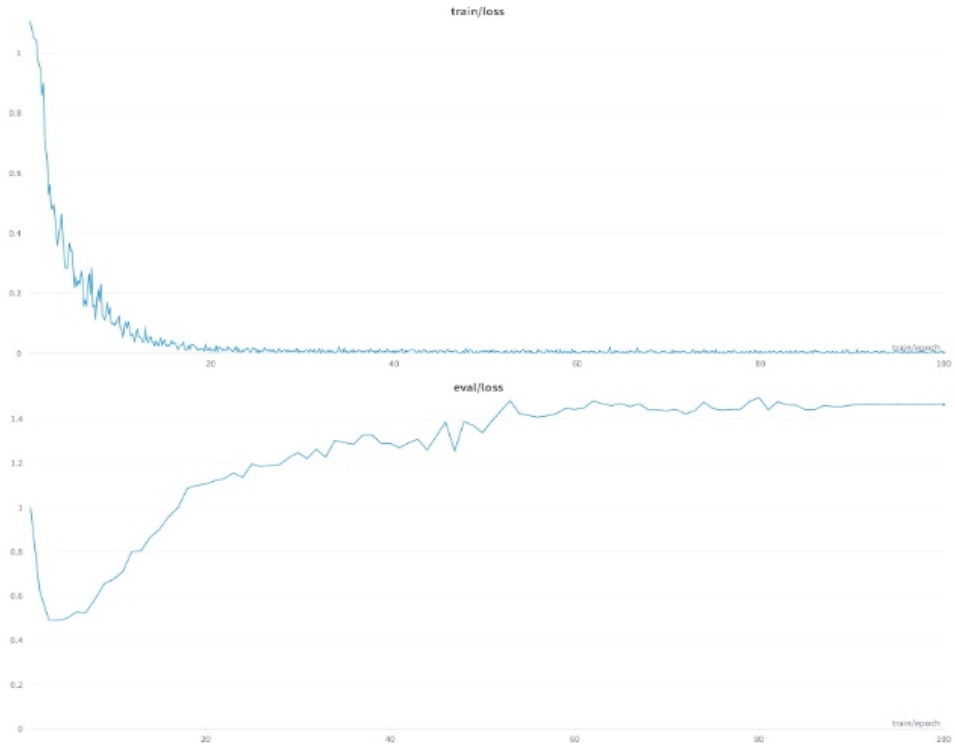


Figure C.2: Graph showing the train and validation loss when fine-tuning on DistilBERT. Looking at the graph, we can understand that with the increase in the number of epochs, the training loss is constantly decreasing and approaching 0. The validation loss decreases till 4 epochs and then starts to increase. Based on this, we fine-tuned DistilBERT for 4 epochs.

Table C.1: Summary of the causal transfer learning results when BERT was used. Here the financial data is used for training, and the gestational diabetes guidelines are used as test data. From the results, we can infer that the F1 score of 0.48 was obtained. With BERT, the F1-score was low for the cause.

	Precision	Recall	F1-Score
E	0.47	0.54	0.51
O	0.70	0.39	0.50
C	0.31	0.63	0.42
Accuracy			0.48
macro avg	0.50	0.52	0.48
weighted avg	0.56	0.48	0.49

Table C.2: Summary of the causal transfer learning results when DistilBERT was used. Here the financial data is used for training, and the gestational diabetes guidelines are used as test data. From the results, we can infer that the F1 score for cause is lower than the other two labels. This result is similar to the BERT's performance, where the F1 score of cause was lower than other labels.

	Precision	Recall	F1-Score
E	0.44	0.68	0.53
O	0.74	0.43	0.54
C	0.35	0.43	0.38
Accuracy			0.51
macro avg	0.51	0.51	0.49
weighted avg	0.57	0.51	0.51

Table C.3: Summary of the causal transfer learning results when LLAMA2 was used. Here the Organizational data is used for training, and the gestational diabetes guidelines are used as test data.

	Precision	Recall	F1-Score
E	0.49	0.53	0.51
O	0.63	0.48	0.54
C	0.34	0.52	0.42
Accuracy			0.50
macro avg	0.49	0.51	0.49
weighted avg	0.53	0.50	0.51