National College of Ireland

Extraction of the Triggering Causes of a Query Event

MSc Research Project
Data Analytics

Srijon Datta
Student ID: 21225265

School of Computing
National College of Ireland

Supervisor: Prof. Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Srijon Datta
Student ID:	21225265
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Prof. Vladimir Milosavljevic
Submission Due Date:	14/08/2023
Project Title:	Extraction of the Triggering Causes of a Query Event
Word Count:	6595
Page Count:	17

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	14th August 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Extraction of the Triggering Causes of a Query Event

Srijon Datta
21225265

Abstract

The main goal of traditional information retrieval systems is to find documents that are pertinent to a certain query idea. But when working with sources like collections of news articles, a user may frequently want to find documents that explain the series of circumstances that may have led to the news event in addition to those that describe the news event itself. Because they involve several underlying causative components, these interactions may be intricate. In response to this demand from the issue, we create the aim of causal information extraction. This work uses a Convolutional Neural Network (CNN) and a Transformer-based model to give an in-depth structure for causality-driven document classification. The overall architecture includes phases for gathering data, extracting information from documents, indexing, and creating input vectors for models. Regarding causal queries, the suggested models successfully separate relevant and irrelevant content. The Transformer-based BERT model outperforms all others in experimental assessment, effectively predicting document relevance with nearly 72% accuracy rate. The work demonstrates the potential of data-driven models in addressing difficult information retrieval problems and identifies prospective directions for model optimization and dataset augmentation in the future.

1 Introduction

It is inherent to human nature to ask "why" and "how" in reaction to every event or circumstance as we attempt to make sense of the situation that we are in. The same is true while seeking to analyze any complex nature of events in current society. As an instance, knowing the 'why' behind the UAE-Israel peace treaty's signing may be useful in analyzing its effects. In order to map events, we typically employ cause and effect relationships. Understanding the relationships between distinct occurrences in cases of cause and effect has long been the goal of the study of cause-and-effect interactions (Asghar; 2016). In certain circumstances, such as the fact that smoking increases the risk of lung cancer, these links are immediately apparent to us. These connections, however, usually entail a combination of a number of causal variables that may have led to the observed event as well as a number of other probable causes that may have, in turn, precipitated events that were already present in these causative factors. In the aforementioned instance, the Israeli settlement plan or Trump's foreign policy might be seen as immediate causal factors (Bowen; 2020). However, if we look more closely at the aforementioned justifications for Israel's settlement policy, we could see some notable components like earning worldwide recognition and mending ties with the Middle East. There are no hard-and-fast rules for how cause-and-effect interactions should be put up

in most situations, according to the literature of Hashimoto et al. (2015) and Riaz and Girju (2014). It can be difficult to explicitly list the causes (in the form of concise text passages) given these complex cause-and-effect linkages. Instead, these causal components are scattered over a number of texts. In that sense, it could be ideal to provide the user this knowledge and give him the responsibility of coming up with potential explanations on his own. Finding words and phrases that are related across texts and to a user query is the main emphasis of conventional IR search engines. Unfortunately, these techniques might not be enough when a user wishes to identify the elements that led to a specific occurrence. A user could conceive of a simple solution in this case, such as adding the "why," "causal factors," or "main causes," etc., as an additional query word. However, in practice, this approach does not help in the actual finding of causal connections. The distinction between causal relevance and traditional topical relevance has been carefully examined by Datta, Ganguly, Roy, Bonin, Jochim and Mitra (2020). This study makes the point that, despite the possibility of some limited-term overlap between documents that are topically and causally relevant for a query, it is anticipated that the majority of these documents will use a different set of terms to describe the various causes that may be to blame for their effects, which typical search mechanisms cannot handle.

1.1 Research Motivation

In conventional searching systems, the user must express their informational needs as a query for search. These questions may occasionally be a direct representation of the user's needs. Human-generated searches, on the other hand, are frequently not very detailed in terms of the user's search objective. An Information retrieval (IR) system retrieves the top-most comparable documents according to the input query, where the level of similarity between the document and the query is determined using a retrieval model's underpinning scoring function, such as BM25, LM-JM, etc.(Hiemstra (2001),Robertson et al. (2009)). The user normally looks through the documents that were returned to find any instances that meet their information needs. It is significant to highlight that while retrieving results from searches for a user, conventional IR systems fail to take the user's search intention into consideration. Finding causal relationships in textual entailment, when cause-and-effect relationships retain distinct conjunctives (such as "because of," "directs to," and so on), has been the focus of several studies (Asghar; 2016). On the other hand, we are interested in situations where cause-and-effect relationships are fairly complicated and lack of any obvious correlations, as is typical of news items (please refer section 1). A key thing to keep in mind while reading news items is that, rather than being straightforward, the reasons of an occurrence are sometimes obscure (Datta, Ganguly, Roy, Bonin, Jochim and Mitra; 2020). Additionally, an event is frequently the result of a number of factors that are spread out across a long period of time. Therefore, it is frequently challenging to locate news reports that would "solely identify" the cause of an occurrence as being one particular event in the past. Because such information is not explicitly reported in news articles, making the initial query more specific by including cause-related keywords, such as "Russia-Ukraine war causes" or "Russia-Ukraine war reasons," etc., and then employing a conventional IR system is not likely to retrieve pertinent information. The hidden associations between the words in various texts can be analyzed in order to uncover such information. As a result, the user of a conventional IR system must use a great deal of time reformulating searches in order to get causally pertinent documents. This is due to the fact that no retrieval engine has been developed

yet that specifically targets the goal of obtaining this specific type of causal information. This study makes the point that, despite the possibility of some restricted term overlap between documents that are topically and causally relevant for a query, it is anticipated that the majority of these documents will use a different set of terms to describe the various causes that may be to blame for their effects, which typical search mechanisms cannot handle. By addressing the following research question, we will attempt to examine this gap in the IR literature in this document.

1.2 Research Question

Given a causal query (rather any effect mentioned in the query), how well a supervised retrieval set up can enumerate the list of plausible triggering causes embedded in the documents?

In order to answer this research issue, we first provide a complete overview of the existing causality research efforts in Section 2. We outline an end-to-end approach that was used throughout the study in Section 3 and the thorough model architecture in Section 4, where each and every phase is covered. Finally, we briefly review our deployed models and their performance in our study in Sections 5 & 6.

2 Related Work

There are numerous ways to investigate the inherent character of cause-and-effect connections from text within the context of textual entailment (Blanco et al.; 2008). However, we are more focused on gathering causal data at the document level rather than concentrating on sentences level. In this part, we provide a high-level assessment of numerous existing approaches developed to discover cause-effect connections in order to contextualize the problem of causal information extraction.

2.1 Identifying the Cause and Effect

Deep neural networks are becoming more and more popular as a result of the increased emphasis on counterfactual (i.e., what may have potentially happened?) causality studies. But initially, establishing the semantic connections between a cause and an effect has been more closely linked to causality (Riaz and Girju (2014); Tanaka et al. (2012)). The idea of using event pairs was eventually inspired by research on the causal linkages between two inquiries conducted by a number of academics (Sun et al.; 2007). This is not the sole attempt to examine causal relationships between two questions, despite the fact that sentence-level effects are typically utilized to capture causal features (Inui and Okumura; 2005). Do et al. (2011) next attempted to show event causality, or the causation of event pairs, inside texts by foreshadowing it (for instance, "police have arrested him" as "he stabbed someone"). The difference between our approach and previous methodologies is that we examine causality that spans a document collection tied to a certain query.

2.2 Techniques Based on Graphs

Graphs make it simple to see how things are connected. When it came to tracing causal inferences, Pearl (2022) recommended using non-parametric charts. Subsequent study by Dawid (2010) used directed acyclic graphs to show causal links before the focus eventually

shifted to Bayesian Networks (Zhang; 2008). Rink et al. (2010), on the other hand, focused on interpreting graph-encoded event- pair causality linkages in text. The main objective of graph pattern-based techniques is to extract event pairs from text and use stochastic measures to study their patterns. Selecting relevant factors from a larger range of events that are likely associated to the query event is the key challenge for our assignment since causal events may not directly relate to the query.

2.3 Causality Knowledge Bases

Since the late 1990s and up until the present, causal research has made use of domain-independent data. Researchers have attempted to fully use the semantic feature of predicate statements (Hashimoto et al.; 2012), which effectively discovers contradicting pairings (e.g., "destroy cancer" $\perp$ "develop cancer"), as well as automated causal connection development (Kaplan and Berry-Rogghe; 1991). These initiatives happened when knowledge-based causality started to take shape. Zhao et al. (2017)'s relational embedding approach emerged after a network of causes and effects was initially built using a set of patterns, expanding the knowledge-base pattern methodology.

2.4 Grouping of the Documents

The importance of causality has also been shown by the classification of documents, where there is typically a complex relationship between features and classes. Paul (2017) developed an approach for detecting important characteristics known as propensity score matching technique with the aim of understanding "which word features cause documents to have the class labels that they do?". Wood-Doughty et al. (2018) treated the causal inference problem as a classification problem and showed how to use logistic regression to look into causality in diverse datasets. The authors took into account characteristics including incomplete records and measurement errors, which commonly obstruct subsequent causal studies.

2.5 Future Event Forecasting

According to Radinsky et al. (2012), new event prediction falls within the category of contingency discourse concerns in NLP, which makes establishing causal links in textual data particularly challenging. Radinsky and Horvitz (2013) began their research by automating the collecting and generalization of a number of events from various web corpora. However, some studies claim that in order to handle causality, there either needs to be two events in the succeeding sentences that have an inter-sentential reliant connection (Riaz and Girju; 2010) or previously trained event-causality linking databases made from web data (Hashimoto et al.; 2014). Therefore, past event knowledge is necessary for future scenario forecasting problems, which is unlikely in our situation since users might not be familiar with the potential root causes of the query event beforehand.

2.6 Method for Responding to Queries

The literature on Natural Language Processing highlights how question-answering systems make use of the essential structure of causality by illuminating the broad scope of causal interactions (Girju; 2003), which aids in locating inter- and intra-sentential causal

connections between words and clauses in response to why" questions (Oh et al.; 2013). Kıcıman and Thelin (2018) have proposed a support system to predict the answers to issues like "Should I join the military?" or "Should I move to California?" The focus of Roemmele et al. (2011) and Gordon et al. (2011) was on common sense causality detection, a unique form of quality assurance. This causality variant helped to resolve ambiguities in discourse linkages and reasoning based on phrase proximity by employing knowledge bases. Because of this, QA techniques either produce semantic or syntactic correlations or extract morphological traits between cause and effect. This is not applicable to operations if an obvious connection between the causative documents and the query event is improbable.

2.7 CNN and Causality Detection Correlation

From 2018 (Narendra et al.; 2018), causality has been incorporated into conventional CNN models. It has also been used to offer general abstractions for deep unsupervised learning methods (Raina et al.; 2009). Harradon et al. (2018) focused on the pertinent ideas deriving from a CNN network in order to estimate the data obtained by activation functions in the network of interest. A knowledge-based CNN, on the other hand, has been proposed by Li and Mao (2019) to identify causal linkages in texts that employ natural language. This body of literature provides a compelling justification for using the CNN model for our study on causation determination.

3 Methodology

Cross Industry Standard Method for Data Mining (CRISM-DM) has been employed in this study. A common approach for carrying out data mining and predictive analytics is CRISP-DM. It offers a methodical and thorough strategy to direct data mining initiatives from inception through implementation. In relation to this study context, the six sequential phases of CRISM-DM have been addressed below.

3.1 Business Understanding

Traditional Information Retrieval systems are not able to deal with the causal relevance of a query event that has various triggering plausible causes behind it. For instance, if someone is interested to find the triggering causes of the Russian invasion of Ukraine, a normal search engine would take the query keywords and fetch those documents where term overlapping can be found. But with this, the actual causes for that particular query event can not be known unless the user digs deeper by themselves to find a satisfactory answer. This research aims to address this problem using supervised techniques so that it can be useful for future search engine optimization.

3.2 Data Understanding

The idea of causation should have a distinct essence from its topical equivalent. As a result, a dataset has been selected for this research in such a way that is specifically designed to capture the causal idea (Datta, Ganguly, Roy, Greene, Jochim and Bonin; 2020), (Datta et al.; 2021).

1. **Target Collection:** FIRE[1] (Palchowdhury et al.; 2011) English Ad-hoc IR collection has been employed for this study. This collection consists of crawling news stories from 'Telegraph India'[2] that were published from 2001 to 2010 over a ten-year span.

2. **Topic Set:** The dataset (Datta et al.; 2021) contains 25 possible causal queries in total. By the term 'Causal Query' we denote those specific type of queries which are related to causal relationships. These types of inquiries are intended to investigate and comprehend the relationships that exist between various variables or incidents. Among 25 queries, 20 randomly picked queries have been used in training the models and the rest of 5 queries have been used for testing purpose using 5-fold cross-validation i.e., the dataset's 25 causal queries will be split into five groups, and the model will go through five iterations of training and testing. Each run will employ a separate set of causal questions as the testing set and the other sets as the training sets. With this method, the model's versatility in responding to different causal inquiries can be robustly evaluated.

3. **Relevance Judgement Set:** Building a document pool for manual relevance evaluation in causal retrieval is more difficult than in conventional information retrieval (IR) for two key reasons. Since there is no recognized paradigm for causal IR, unlike standard IR, it might be difficult to incorporate pertinent information. In order to effectively analyze causal links, assessors also require prior knowledge of the event indicated in the inquiry. A multi-query formulation exploratory technique was employed to overcome this. During investigation, an interactive system assisted bookmark papers, highlighting those that could be relevant for developing causal relationships. These bookmarked documents were gathered into an evaluation pool together with the top 100 documents that were located using conventional IR models (e.g., LM, BM25, RLM). To establish the document's relevance, assessors made binary decisions based on their existing knowledge and the findings of their explorations (Datta, Ganguly, Roy, Bonin, Jochim and Mitra; 2020).

3.3 Data Preparation

1. **XML Parsing and Extracting the Raw Texts:** The Dataset which has been used in this study is the collection of crawling news stories from an Indian newspaper, Telegraph India, that were published from 2001 to 2011 over ten years. These news articles are categorized into different topics namely 'frontpage', 'nation', 'national', 'sports', 'bengal' etc. Each of the category contains a list of news articles where each article contains two tags namely '<DOCNO>' and '<TEXT>' where the prior one hold a unique document id for each of the article and the later one holds the news description. To make a single raw text file from all the news articles, XML parser has been used.

   Firstly, two tags have been identified and the contents are stored in new variables after removing all kinds of punctuation, special characters, and excess blank spaces using the 'Regex' library in Java (Refer to the 'MakeTelegraphDump.java' code file).

[1] `http://fire.irsi.res.in/fire/static/data`

[2] `https://www.telegraphindia.com/`

Secondly, the parsed contents have been written in the file named 'telegraph_01_11. dump' (Refer to line number 44 of the same code file).

Thirdly, stemming and stop words removal process has been employed for this parsed collection so that the raw texts can be obtained and all the words can be converted into numerical vectors in a multi-dimensional space. This process, commonly known as word embedding, has been done using the "Word2Vec" technique. The query dataset, consists of 25 queries, has been paesed in the similar fashion as described above using xml parser to get the raw topics.

2. **Indexing the Collection:** To index the whole dataset collection, *Lucene 8.8* (*Apache Lucene*; 2023), a Java API, has been used in this work. Lucene is a popular Java-based open-source information retrieval library. It offers resources for effectively indexing and looking up text-based material. The process of converting text materials into a structure that enables quick and efficient searches is known as lucene indexing. This is essential because it converts unstructured textual information into a structured format that makes it easier to conduct quick, precise, and sophisticated searches.

   Firstly, each document text has broken into smaller units called tokens. Tokenization entails breaking the text down into individual words, phrases, or terms.

   Secondly, to guarantee uniformity, tokens have been normalized which include stemming (i.e. reducing words to their root or basic form), or changing all letters to lowercase. The normalizing enhances search precision.

   Thirdly, Tokens are linked to the documents that include them using an inverted index data structure in lucene. The documents' unique words are included in an inverted index alongside a list of the document numbers that correspond to each use of the term.

   Forthly, for each of the indexed document, a document vector has been produced using lucene. The document's unique identity, creation date, and other information are all included in this vector of document metadata. The terms in the document and their locations inside are also mentioned in the document vector.

   Finally, Lucene builds the document vectors and updates the inverted index during the indexing process after reading each document's tokenized and normalized content. The key benefit of utilizing lucene is that we can run searches on the indexed data when the indexing process is complete. Lucene uses the inverted index to swiftly find documents that contain the search phrases when we do a query search. Using methods like the TF-IDF (Term Frequency-Inverse Document Frequency) model, it ranks the results according to relevancy.

3.4 Modelling

The prime objective for this piece of work had always been to employ supervised learning techniques to find the relevant causes for a causal query. Keeping the research objective in mind and based on the existing related works, Convolutional Neural Network (CNN) for document classification and Bidirectional Encoder Representations from Transformers (BERT) have deployed in this work. In the Section 5, a detailed descriptions about both the models (CNN+BERT) has been given including the model architectures.

3.5 Evaluation

Generally speaking, the best way to evaluate any classification problem is to measure how accurate is the model outcome, which in other words predicts the no. of correctly predicted relevant/non-relevant documents compared to the true relevance of the same set of documents. We report the measure on a scale of 100.

3.6 Deployment

The researcher shows how the whole framework actually functions in real-time during the deployment phase of the study. For given causal query which as an event mentioned in it, our suggested model is able to identify the those documents which are potential causes for that particular event.

4 Design Specification

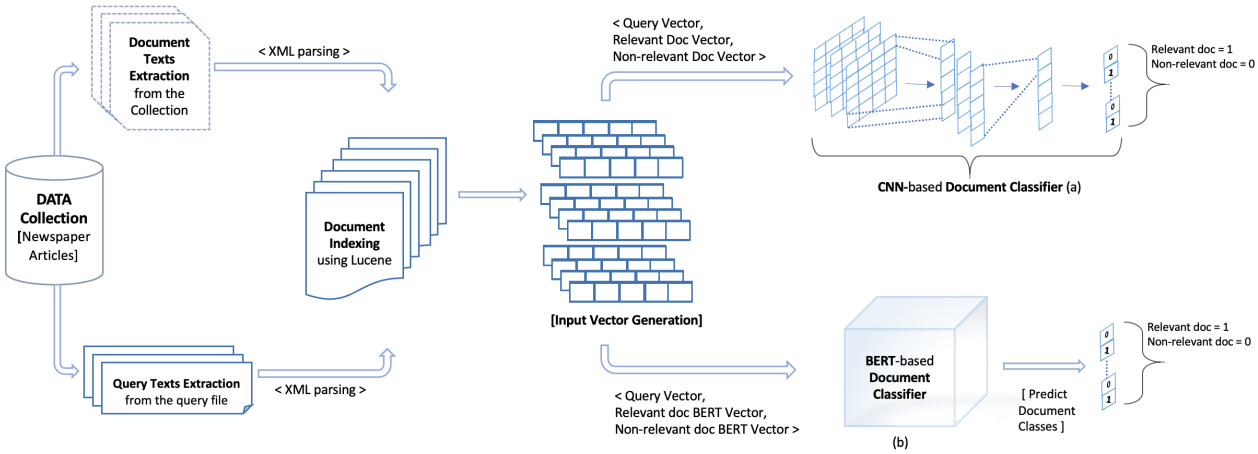

Figure 1: Design Specification

Figure 1 depicts the full design architecture which has been followed in this study. Starting from the Data Collection, through Document & Query Text Extraction, indexing the parsed documents (using lucene) which have been used to generate Input Vectors for the models (detailed explanation has been given in the Section 5) and deployment of the models to get the final relevant/non-relevant classification of the documents, all have been incorporated in one diagram. Nevertheless, in Section 5, more elaborated model architectures have been given for both the models (CNN & BERT).

5 Implementation

In this section, we explain the end-to-end causality-based classification model implementation in detail. As mentioned in Section 4, we proposed two different models, 1. CNN-

based model Afzal et al. (2015) and 2. Transformer-based model (specifically BERT) Devlin et al. (2018).

5.1 Environment setup

Before explaining the implementation details of our proposed two models, we first enumerate the list of dependencies in terms of setting the respective environment. It is worth noting here that the proposed CNN-based model is comprised of four different modules, such as, (a) collection indexing, (b) input vectors generation (i.e. query-document interaction matrices), (c) training phase and (d) testing phase. Whereas, the transformer-based model includes three distinct steps, such as, (1) BERT vectors generation for model input, (2) training phase and (3) testing phase.

In terms of model implementation, collection indexing and input vector generation for CNN model are developed using Java and rest of the modules both for CNN and BERT-based models are developed using Python. Each of the modules makes use of a number of Java and Python packages as required.

Precisely, a dedicated virtual environment was created to implement aforementioned modules at different stages. For instance, for indexing the collection and generating input vectors for CNN-based classification model, we need the following java packages.

- `JDK 1.8.0 or above`
- `lucene 5.3.1`
- `archive-commons-1.12.0.jar`
- `commons-compress-1.12.jar`
- `commons-io-2.5.jar`
- `commons-lang-2.3.jar`
- `commons-math3-3.6.1.jar`
- `lucene-analyzers-common-5.3.1.jar`
- `lucene-backward-codecs-5.3.0.jar`
- `lucene-core-5.3.1.jar`
- `lucene-queries-5.3.1.jar`
- `lucene-queryparser-5.3.1.jar`

Figure 2: List of Java packages required.

The modules where CNN and transformer-based architectures are involved, they leverage a couple of python libraries namely the following.

- conda 4.8.2
- python 3.7.9
- numpy 1.19.4
- keras 2.3.0
- tensorflow 2.2.0
- scikit-learn 0.23.2
- nltk 3.5
- transformers 4.6.1
- more_itertools 8.13.0
- numpy 1.21.6
- pyterrier 0.1.5
- python_terrier 0.8.1
- torch 1.10.0
- transformers 4.20.1

Figure 3: List of Python packages required.

5.2 CNN-based model architecture

The first model that we propose is mainly based on extracting term-semantics interaction at two different stages. Firstly, the intra-query level of modelling is done, i.e. the interaction vector is computed between the query and the pseudo-relevant document retrieved by the same query as in there in the qrels. Next, the information is captured at the inter-query level, to model the relative relevance measure.

Query-document interaction The proposed term overlap based encoding is based on the word embedding based interactions in the deep relevance matching model Guo et al. (2016). Rather than applying separate encoders for documents and queries, this method first computes the interaction between a query Q and any document D from the relevance judgements as a fixed length vector by quantizing the cosine similarity values between every term pair – one from the query Q and the other from either the relevant document D_r or a non-relevant document D_{nr}. The quantization step involves the number of intervals that indicates the range of cosine similarity values ($[-1, 1]$) is partitioned.

Causal CNN architecture In the next step, a 2D convolutional neural network is employed that takes as input a $k \times N \times p$ dimensional interaction tensor – one for the

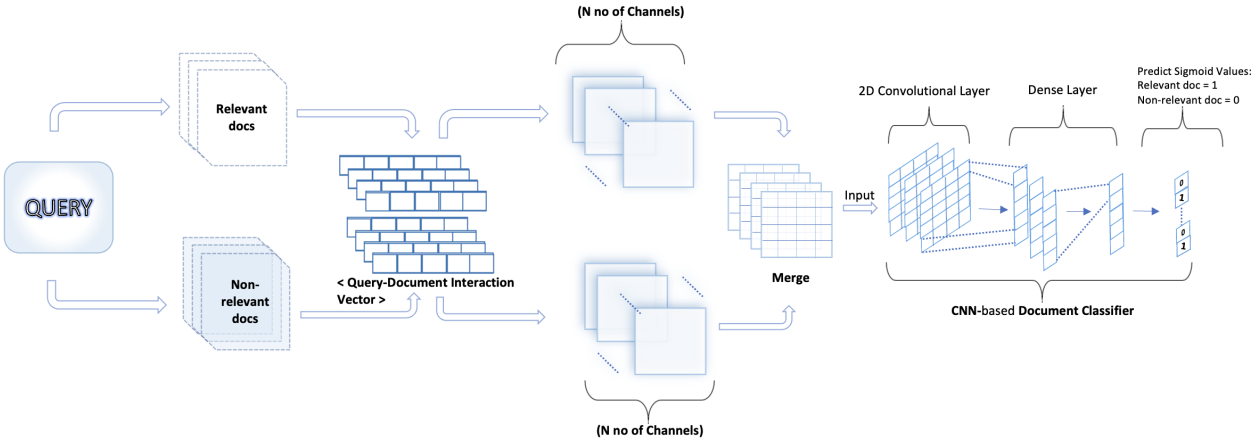

Figure 4: CNN-based causal model architecture

query and the relevant document D and the other for the query and the non-relevant D_{nr} one. This slices the tensor into N channels and transforms each to yield a fixed-length vector after the standard flattening step, following the application of the convolutional filters. The 2D-CNN encoded vectors is then given by a merge operation followed by a dense layer of parameters finally leading to a sigmoid for the binary prediction of the class, i.e. if the given input document is relevant or not. The 2D-CNN coupled with the merge and the dense layers thus define the entire set of parameters in the realisation of the generic model.

5.3 Transformer-based model architecture

The other classification model that we propose is a transformer-based neural model which feeds in BERT-based uncased vectors of both queries and documents. Instead of inputting query-vector semantics, this model takes raw query and documents texts and then compute the 768 dimensional BERT vectors for further feature extraction.

Query-document BERT vectors The 2D-CNN-based encoding makes use of individual word vectors to obtain interaction tensors, which are then supplied as inputs to a neural network. Unlike the idea of early interaction, this model is purely data-driven, where the model computes features from it 768 dimensional BERT-base-uncased vectors. The input query and each relevant and non-relevant documents are first tokenized and thus generates transformer-encoding suitable BERT vectors.

Causal BERT architecture The proposed causal model takes both relevant and non-relevant vectors together and thus classifies them according to the model prediction output. Firstly, the tokenized BERT vectors of a query and a document is passed through a siamese network. The siamese network merge the vectors of both relevant and the non-relevant query-document texts. Features are then captured and passed through a dense layer for training and optimizing the learning parameters. The dense layer is then flatten and passed through a sigmoid layer which predicts the relative relevance of the input documents by binarizing the output sigmoid values and thus classifies the input

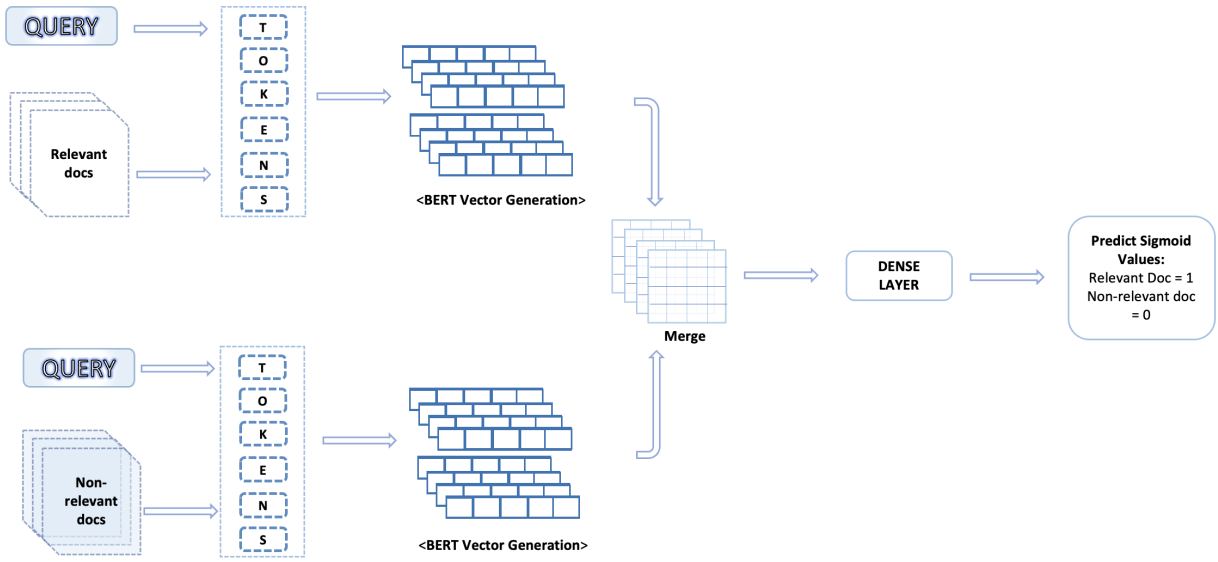

Figure 5: BERT-based causal model architecture

documents as relevant or non-relevant.

6 Evaluation

This section first explains how we prepared the ground truth for model outcome evaluation. Then we illustrate two of our model outcome and what accuracies are achieved and thus which model performs the best for this causal classification task. Lastly, through graphical presentation, we show the how the models achieve accuracies as the number of epochs increases and a comparative analysis of per query prediction accuracy through two different proposed models.

6.1 Ground truth generation

The model must first create all plausible pairs of a relevant document $D_r \in QREL(Q)$ and a non-relevant document $D_{nr} \in QREL(Q)$, where D_r and D_{nr}, respectively, denote a relevant and a non-relevant document present in the relevance judgement set in response to the query Q. This is done for a set of training queries $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$. As a result, for each pair of (D_r, D_{nr}) or (D_{nr}, D_r), the ground truth is prepared so that the pair (D_r, D_{nr}) is labeled as 1 if the relevance score of $(D_r - D_{nr}) > 0$ and 0 otherwise.

6.2 Measuring accuracy

Both the proposed CNN and transformer-based models leverage the advantage of classification. The model's efficacy is then measured in terms of 'accuracy' when it performs as a binary classifier and predicts which of any given two input documents is relevant. Thus, we measure the % of total correctly predicted documents per query compared to the ground truth. Experimental results show that CNN-based model achieved nearly 66%

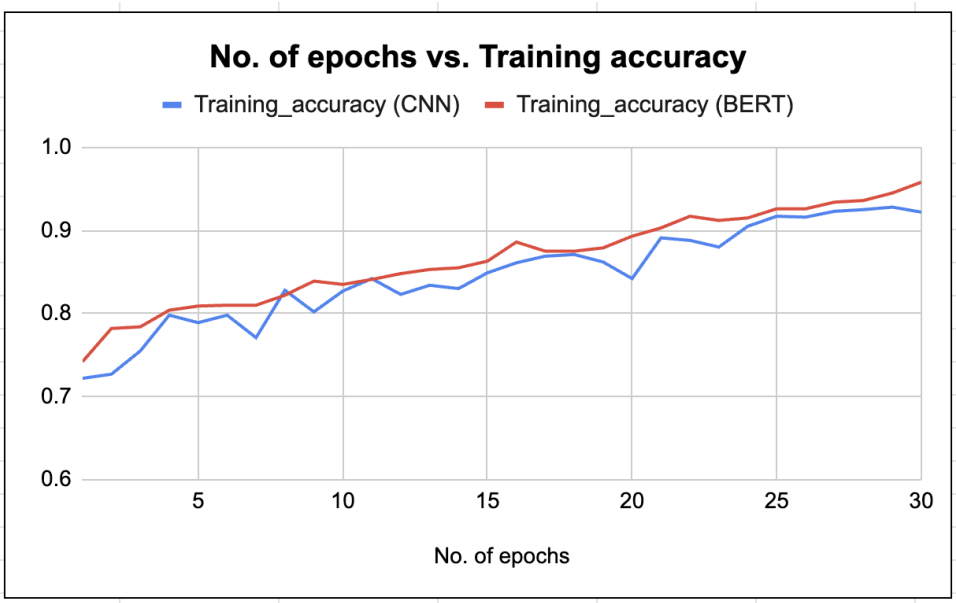

Figure 6: Distribution of Training Accuracies for both the models (CNN+BERT) for different Epochs

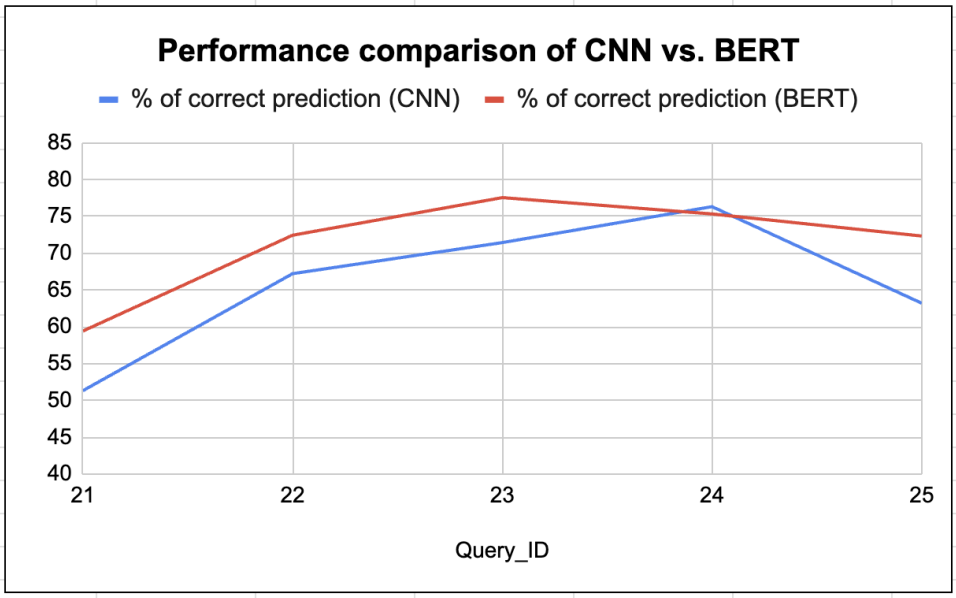

Figure 7: Performance Comparison

accuracy, where, transformer-based model outperformed the former one with nearly 72% accurate predictions.

6.3 Discussion

This thesis compares two of the state-of-the-art neural model architectures, i.e. Convolution NN and transformer based BERT model to solve the causality-driven document classification problem. Extensive experimental results show that the latter performs better than that of CNN-based model. Figure 6 depicts the training accuracy achieved by both the models as the number of epoch increases which emphasizes the fact that BERT

vectors are able to capture more useful features than query-document semantic features. Not only that, it is also evident that transformers are useful to gain consistent In addition to this, we also report the % of correctly predicted document class and from the Figure 7, it is clear that BERT-based model was able to predict maximum no. of correct classes in comparison with CNN.

It is worth mentioning that the proposed models could be improved further at a couple of following points. Firstly, due to brevity of time, models parameters could not be tuned in a large range to obtain the best outcome. Usually, these neural models are time-consuming, therefore, parameter tuning to a greater extent would have been too expensive for timely finish of this thesis.

Another scope of improvement includes the size of the dataset used for experiments. Generally speaking, the neural models that are used in this thesis are data-hungry. The training would have been more accurate and the learning parameters could be more optimized by using a larger collection which is not available at the moment. So all these gaps are considered as our future scope of improvements.

7 Conclusions and Future Work

Faced by any situation in our daily life, it is our basic human instinct to reveal the background story of any event that occurred. This in turn leads us to find the answer behind any *'why'* question. This thesis aims to uncover such a list of potential causally relevant documents that might have led to any event occur, mentioned in a query. In line to this objective, two different models are proposed, one is a feature-based method that computes the term-semantics between query and documents and then classifies causally relevant and non-relevant ones; and the other method is a completely data-driven where the BERT vector of queries and corresponding pseudo-relevant documents are fed into a transformer-based architecture and thus classifies the causal relevant and non-relevant documents.

This thesis illustrates the efficiency of both the models via thorough experiments and analysis on a benchmark causal dataset available hosted by FIRE. The extensive experimental results show that transformer-based purely data driven model outperforms that of feature-based CNN model although at the cost of more time and resource. The transformer-based classifier obtained nearly 72% prediction accuracy, whereas the CNN-based predictor achieved almost 66%.

This thesis formulates and solves the causality-driven information retrieval problem as a classification task by leveraging the state-of-the-art convolution neural model and transformer based model as well. In future, we aim to solve this causal document extraction problem as a neural re-ranking task so that any intended search user might not only be able to uncover if any given document is causally connected or not, but also they would know which causal document has the most relevance.

8 Acknowledgement

I want to thank my supervisor, Prof. Vladimir Milosavljevic, for his advice and assistance with the research project. He has given all the essential information needed to do the assignment properly.

References

Afzal, M. Z., Capobianco, S., Malik, M. I., Marinai, S., Breuel, T. M., Dengel, A. and Liwicki, M. (2015). Deepdocclassifier: Document classification with deep convolutional neural network, *2015 13th international conference on document analysis and recognition (ICDAR)*, IEEE, pp. 1111–1115.

Apache Lucene (2023). `https://lucene.apache.org/`. Accessed: 2023-07-11.

Asghar, N. (2016). Automatic extraction of causal relations from natural language texts: a comprehensive survey, *arXiv preprint arXiv:1605.07895* .

Blanco, E., Castell, N. and Moldovan, D. I. (2008). Causal relation extraction., *Lrec*, Vol. 66, p. 74.

Bowen, J. (2020). Five reasons why israel's peace deals with the uae and bahrain matter, *BBC Middle East editor, 15th September 2020* .
URL: *https://www.bbc.com/news/world-middle-east-54151712*

Datta, S., Ganguly, D., Roy, D., Bonin, F., Jochim, C. and Mitra, M. (2020). Retrieving potential causes from a query event, *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1689–1692.

Datta, S., Ganguly, D., Roy, D. and Greene, D. (2021). Overview of the causality-driven adhoc information retrieval (CAIR) task at FIRE-2021, *FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, India, December 13 - 17, 2021*, ACM, pp. 25–27.

Datta, S., Ganguly, D., Roy, D., Greene, D., Jochim, C. and Bonin, F. (2020). Overview of the causality-driven adhoc information retrieval (CAIR) task at FIRE-2020, *FIRE 2020: Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020*, ACM, pp. 14–17.

Dawid, A. P. (2010). Beware of the dag!, *Causality: objectives and assessment*, PMLR, pp. 59–86.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .

Do, Q., Chan, Y. S. and Roth, D. (2011). Minimally supervised event causality identification, *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 294–303.

Girju, R. (2003). Automatic detection of causal relations for question answering, *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, pp. 76–83.

Gordon, A., Bejan, C. and Sagae, K. (2011). Commonsense causal reasoning using millions of personal stories, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 25, pp. 1180–1185.

Guo, J., Fan, Y., Ai, Q. and Croft, W. B. (2016). A deep relevance matching model for ad-hoc retrieval, *Proc. 25th ACM CIKM*, CIKM '16, Association for Computing Machinery, New York, NY, USA, p. 55–64.

Harradon, M., Druce, J. and Ruttenberg, B. (2018). Causal learning and explanation of deep neural networks via autoencoded activations, *arXiv preprint arXiv:1802.00541* .

Hashimoto, C., Torisawa, K., De Saeger, S., Oh, J.-H. et al. (2012). Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 619–630.

Hashimoto, C., Torisawa, K., Kloetzer, J. and Oh, J.-H. (2015). Generating event causality hypotheses through semantic relations, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29, pp. 2396–2403.

Hashimoto, C., Torisawa, K., Kloetzer, J., Sano, M., Varga, I., Oh, J.-H. and Kidawara, Y. (2014). Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 987–997.

Hiemstra, D. (2001). *Using language models for information retrieval*, Citeseer.

Inui, T. and Okumura, M. (2005). Investigating the characteristics of causal relations in japanese text, *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pp. 37–44.

Kaplan, R. M. and Berry-Rogghe, G. (1991). Knowledge-based acquisition of causal relationships in text, *Knowledge Acquisition* **3**(3): 317–337.

Kıcıman, E. and Thelin, J. (2018). Answering what if, should i and other expectation exploration queries using causal inference over longitudinal data, *Conference on Design of Experimental Search and Information Retrieval Systems (DESIRES)*.

Li, P. and Mao, K. (2019). Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts, *Expert Systems with Applications* **115**: 512–523.

Narendra, T., Sankaran, A., Vijaykeerthy, D. and Mani, S. (2018). Explaining deep learning models using causal inference, *arXiv preprint arXiv:1811.04376* .

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

Palchowdhury, S., Majumder, P., Pal, D., Bandyopadhyay, A. and Mitra, M. (2011). Overview of FIRE 2011, *Multilingual Information Access in South Asian Languages - Second International Workshop, FIRE 2010, Gandhinagar, India, February 19-21, 2010 and Third International Workshop, FIRE 2011, Bombay, India, December 2-4, 2011, Revised Selected Papers*, pp. 1–12.

Paul, M. (2017). Feature selection as causal inference: Experiments with text classification, *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 163–172.

Radinsky, K., Davidovich, S. and Markovitch, S. (2012). Learning causality for news events prediction, *Proceedings of the 21st international conference on World Wide Web*, pp. 909–918.

Radinsky, K. and Horvitz, E. (2013). Mining the web to predict future events, *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 255–264.

Raina, R., Madhavan, A. and Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors, *Proceedings of the 26th annual international conference on machine learning*, pp. 873–880.

Riaz, M. and Girju, R. (2010). Another look at causality: Discovering scenario-specific contingency relationships with no supervision, *2010 IEEE Fourth International Conference on Semantic Computing*, IEEE, pp. 361–368.

Riaz, M. and Girju, R. (2014). In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs, *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 161–170.

Rink, B., Bejan, C. A. and Harabagiu, S. M. (2010). Learning textual graph patterns to detect causal event relations., *FLAIRS Conference*.

Robertson, S., Zaragoza, H. et al. (2009). The probabilistic relevance framework: Bm25 and beyond, *Foundations and Trends® in Information Retrieval* **3**(4): 333–389.

Roemmele, M., Bejan, C. A. and Gordon, A. S. (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning., *AAAI spring symposium: logical formalizations of commonsense reasoning*, pp. 90–95.

Sun, Y., Xie, K., Liu, N., Yan, S., Zhang, B. and Chen, Z. (2007). Causal relation of queries from temporal logs, *Proceedings of the 16th international conference on World Wide Web*, pp. 1141–1142.

Tanaka, S., Okazaki, N. and Ishizuka, M. (2012). Acquiring and generalizing causal inference rules from deverbal noun constructions, *Proceedings of COLING 2012: Posters*, pp. 1209–1218.

Wood-Doughty, Z., Shpitser, I. and Dredze, M. (2018). Challenges of using text classifiers for causal inference, *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, Vol. 2018, NIH Public Access, p. 4586.

Zhang, J. (2008). Causal reasoning with ancestral graphs, *Journal of Machine Learning Research* **9**: 1437–1474.

Zhao, S., Wang, Q., Massung, S., Qin, B., Liu, T., Wang, B. and Zhai, C. (2017). Constructing and embedding abstract event causality networks from text snippets, *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 335–344.