Answering Binary Causal Questions: A Transfer Learning Based Approach

Humayun Kayesh[†], Md. Saiful Islam[†], Junhu Wang[†], Shikha Anirban[†], A.S.M. Kayes[‡], Paul Watters[‡]
[†]*School of Information and Communication Technology*
Griffith University
Gold Coast, Australia
[‡]*School of Engineering and Mathematical Sciences*
La Trobe University
Melbourne, Australia
{h.kayesh,saiful.islam,j.wang}@griffith.edu.au, shikha.anirban@griffithuni.edu.au,{a.kayes, p.watters}@latrobe.edu.au

***Abstract*—Causal question answering is a task of answering causality related questions. The questions are referred to as binary causal questions when the questions e.g., "*Could X cause Y?*" can be answered by *yes/no* answers. Answer to the previous question is *yes* if X is a cause of Y, and otherwise *no*. The binary causal question answering systems can be used to validate causal relationships, which can be particularly useful for decision making. For example, it could be useful for the tourism authorities to know the answer to the question "Could *growing social tension* cause *reduction in tourism*?". We aim to automatically answer such binary causal questions by developing a machine learning model. However, training a machine learning model to detect causal relationships is challenging due to the lack of large and high quality labeled datasets. In this paper, we propose a transfer learning-based approach which fine-tunes pretrained transformer based language models on a small dataset of cause-effect pairs to detect causality and answer binary causal questions. The proposed approach achieves performance comparable to a number of benchmark approaches on five benchmark test datasets extracted by human experts conditioned on the same small training dataset.**

I. Introduction

Binary causal questions ask whether there is a relationship between a candidate cause and a candidate effect. For example, in the following question "Could *Australian bushfire* cause *a jump in carbon concentrations in the atmosphere*?"[1], "*Australian bushfire*" is a candidate cause and "*a jump in carbon concentrations in the atmosphere*" is a candidate effect. If there exists a causal relationship between the candidate cause and the candidate effect, the answer to the above binary causal question is either *yes* or *no*. The automatic answering of such binary casual questions might play an important role in everyday decision making and reasoning. It can be explored as an important tool by the decision makers to assess the situation and take informed decision. Additionally, a binary causal questions answering model may be used to discover new and previously unknown pairs of cause and effect as exemplified and illustrated in Fig. 1 for "Could Wuhan coronavius cause pneumonia?"[2] and "Could climate change cause arson?"[3].

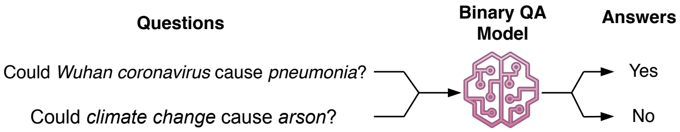

Questions
Could *Wuhan coronavirus* cause *pneumonia*?
Could *climate change* cause *arson*?
Binary QA Model
Answers
Yes
No

Fig. 1: An illustrative example of binary causal question answering (QA) model

Answering causal questions is a trivial skill of human beings and often a human can answer causal questions without much effort as humans acquires knowledge about causality from the very beginning of their life through day-to-day life experiences and continuous learning. However, this trivial task for human is quite challenging for computers. To answer binary causal questions automatically, appropriate background knowledge (labeled training dataset) on causality is crucial, which is often hard to achieve. In the literature, there exists a number of approaches [1]–[11] to solve the automatic causality detection problem. The approaches apply various techniques such as linguistic rules [1], association rule mining [2], [3], causal network analysis [4], [5], and recently, application of neural network-based algorithms [6]–[11]. The rule-based approaches are often domain specific and not generalizable. The causal network-based approaches rely on exact word matching and neural network-based approaches require large training datasets. In general, the supervised machine learning-based approaches for causality detection require large and high quality labeled datasets to train. Unfortunately, high quality training data is often either expensive to acquire or not available. Also, when the training dataset becomes large, training machine learning models on such large datasets be-

[1] "The devastating bushfires in Australia are likely to cause a jump in carbon concentrations in the atmosphere this year, a forecast suggests, bringing the world closer to 1.5C of global heating." - https://www.theguardian.com/environment/2020/jan/24/australian-bushfires-will-cause-jump-in-co2-in-atmosphere-say-scientists

[2] "So far, we know the new Wuhan coronavirus causes pneumonia and therefore places an extra burden on hospitals."-http://theconversation.com/should-we-be-worried-about-the-new-wuhan-coronavirus-130366

[3] "Victoria police say there is no evidence any of the devastating bushfires in the state were caused by arson."-https://www.theguardian.com/australia-news/2020/jan/08/police-contradict-claims-spread-online-exaggerating-arsons-role-in-australian-bushfires

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

comes computationally expensive. These drawbacks limit the applicability of such approaches. Hence, a machine learning approach that requires relatively small dataset to train, but still contains enough background knowledge on causality and achieves reasonably comparable performance is desirable.

In this paper, we aim to answer binary causal questions by applying transfer learning based approach as it allows us to utilize existing generalized background knowledge for our task specific needs by fine-tuning the model with a relatively small training dataset. We apply pretrained transformer based language models such as Bidirectional Encoder Representation from Transformers (BERT) and its variants to exploit the already learned model weights to solve the problem of large dataset requirements. In particular, we found that the *WordPiece* tokenization [12] technique used in BERT models is effective to deal with the exact word matching problem in our task. We also proposed an approach to automatically extract a training dataset consisting of cause and effect pairs from news articles by applying causal cue words and then use this dataset to fine-tune the models for our binary causal question answering task. We list our contributions as given as below:

1) we propose a transfer learning-based approach that applies BERT language models to answer binary causal questions;
2) we also propose an automatic approach to collect quality training dataset from news articles to fine-tune the BERT models to detect causality; and
3) we perform extensive experiments to show the effectiveness of the proposed approach.

The rest of the paper is organized as follows. Section II reviews the existing state-of-the-art approaches. Section III presents our approach. Section IV discusses the experimental results. Finally, Section V concludes the paper.

II. Related Work

Causality detection from text and question answering are both active research areas. The causality detection approaches that are relevant to our problem includes linguistic rules-based approach, graph-based approach, machine learning-based approach and very recently neural network-based approach.

A rule-based approach on causality detection is proposed by Radinsky et al. [1] that automatically generates a set of rules that are applied to detect causal relationship. However, the rules are dependent on the grammatical correctness of sentences, which is not always available or sometimes missing in open texts. Background knowledge is often useful to detect causality. The graph-based approaches [4], [5] encode background knowledge in causal network and use that knowledge to detect causality. Luo et al. [4] propose a graph-based approach that collects a set of causal sentences from web using a set of causal cue words. Sentences are then split into causal phrases and effect phrases. A Cartesian product between the words in a causal phrase and the corresponding effect phrase is taken to prepare a list of word pairs. The first word in the pair comes from the causal phrase and the other word comes from the effect phrase. The pairs are then used to build a directed causal graph. The nodes in the graph correspond to the words and the edges represent a causal relation. Each edge contains a weight. For example, a directed link from node A to node B represents the number of times the word A were used as a cause and B as an effect. These values are then used to calculate the combined causal scores of candidate cause-effect pairs. However, this approach cannot capture the multi-word expression as it tokenizes every word in a phrase [5]. Also, this approach cannot detect causality if either of the words on a pair is missing in the causal graph.

A causal questions answering method is proposed in [8] that aims to rerank answers to the causal questions using embeddings. This approach collects causal pairs from free text and then converts them into causal embeddings. Both causal and effect embeddings are then passed to two separate Convolutional Neural Networks (CNN). The outcomes of the CNNs are Max Pooled and then merged together by calculating cosine similarity. The *Softplus* activation function is used in the final layer with a single node to detect causality. Since this approach is trained solely on the training data, it requires a large training data for training. Dasgupta et al. [6] propose a neural network-based approach to extract cause, effect and causal connectives in a sentences. The approach uses both word features and linguistic features to train a Bidirectional Long Short-Term Memory (Bi-LSTM) model to detect causality. The input to the model is a combination of word vectors and linguistic vectors. The authors used a pretrained GloVe model to convert each word into a 300-dimensional vector. The linguistic features are composed of both syntactic features and semantic featuers. The syntactic featuers include part-of-speech tags, dependency relationship, and positions of noun and verbs. The semantic features includes nine noun hierarchy in WordNet proposed in [13] along with the grammatical structure of the sentence. Finally, the trained model is used to label each word whether the word is a cause, an effect, a causal connective or none. However, none of the above approaches targets automatic binary causal question answering task.

To the best of our knowledge, there exists one work on binary causal question answering task and the approach is called Natural Language Model - Bidirectional Encoder Representation from Transformers (NLM-BERT), which is proposed by Hassanzadeh et al. [7]. The authors assume that the relationship between cause and effect is unidirectional. More specifically, if (X, Y) is a causal pairs and “X may cause Y” is true, then it is less likely that “Y may cause X” will be true. The authors collected a corpus of 17 million (17M) causal sentences and converted them into vectors using a pretrained BERT model. For a candidate (X, Y) pairs, it finds 10 closest sentences (k-nearest neighbor approach) for both “X may cause Y” and “Y may cause X”. Then it calculates the average cosine similarity for both of the cases. Two threshold values are used to finally derive the yes/no answer for “Could X cause Y?”. The main drawback of this approach is that the threshold values are not previously known and varies from test dataset to another test dataset. Also, it requires expert

TABLE I: List of causal cue words used to prepare the training dataset (adapted from [9], [10])

affect	because	causes	due to	if	induce	owing to	results from
affected by	because of	causing	effect of	if..., then	induced	reason for	so that
affects	bring on	consequently	for this reason alone	in consequence of	inducing	reason of	that's why
and consequently	brings on	coz	gave rise to	in response to	lead to	reasons for	the result is
and hence	brought on	coz of	give rise to	inasmuch as	leading to	reasons of	thereby
as a consequence	cause	decrease	given rise to	increase	leads to	result from	therefor
as a consequence of	caused	decreased by	giving rise to	increased by	led to	resulted from	thus
as a result of	caused by	decreases	hence	increases	on account of	resulting from	

knowledge to set the correct threshold values.

In this paper, we propose a transfer learning-based approach to deal with the above mentioned challenges of automatic answering binary causal questions. We propose to use a pretrained BERT model similar to Hassanzadeh et al. [7] that comes with a rich linguistic information. Unlike the approach proposed by Hassanzadeh et al. that applies k-nearest neighbour-based cosine similarity calculation on large dataset, we fine-tune the BERT model with a relatively small training dataset to contextualize the model for the binary causal question answering task. In our approach, we also avoid the manual threshold selection for the test datasets - threshold selection is not dependent on test dataset. Once our model is fine-tuned on the training dataset, it can be applied for each of the test datasets without the need of selecting any dataset specific thresholds.

III. Our Approach

In this section, we describe our transfer learning-based approach to answer binary causal questions. At first, we extract and prepare a training dataset from news articles and then we fine-tune BERT models on this training dataset. The fine-tuned BERT model is then used to answer binary causal questions.

A. Extraction of Training Data

We apply a semi-supervised approach to prepare the training dataset. In our approach, we extract a set of causal and non-causal pairs from one million news articles [14]. To get the causal pairs, we extract the sentences that contain at least one causal cue words e.g., *causes, due to, and because of* (see the complete list as given in Table I). Then we extract causal pairs such as (X, Y) from the sentences using those causal cue words where X corresponds to the causal phrase and Y corresponds to the effect phrase. In our training dataset, we label each of these pairs as *causal*. We also extract non-causal training data from the same news articles dataset. To avoid any data overlap with the causal pairs, we extract the sentences that do not contain any causal cue words. We divide each sentence into half to prepare the non-causal pairs (X, Y), where X represents the first half and and Y represents the second half of the sentence. Each of these pairs of data is labeled as *not_causal* as we assume that these sentences should not contain any causality.

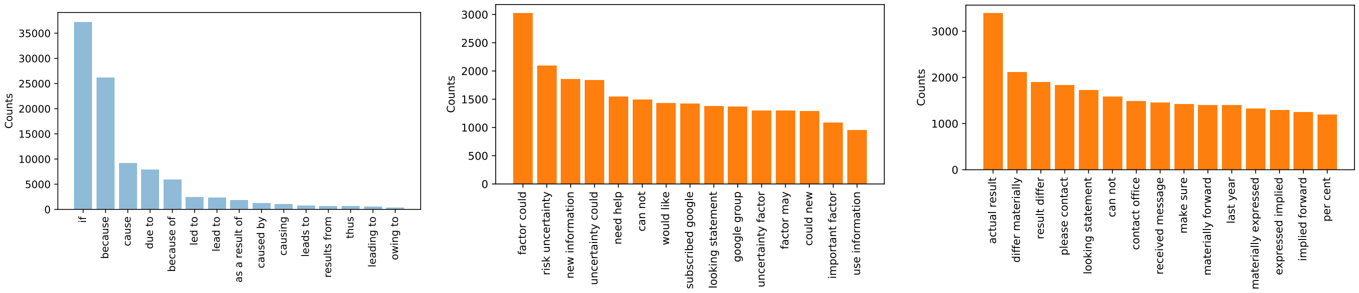
TABLE II: Training dataset statistics

Name	Causal	Non_causal
Number of pairs	100000	100000
Vocabulary size	27206	63471
Longest phrase size	55	189
Shortest phrase size	1	2
Average phrase size	3.5	9.23

Table II illustrates the training dataset statistics which includes the total number of pairs, vocabulary size and the size of the phrases. One prominent characteristic that is visible from the table is that the non-causal pairs in the training dataset are around 3 times longer than the causal pairs. Also, the non-causal pairs contains more unique vocabulary than the causal pairs, i.e., there are 27206 and 63474 unique words in the causal and non-causal pairs, respectively. This is because, we applied linguistic rules such as "X causes Y" and "Y because of X", to extract the causal phrase X and the effect phrase Y from the sentence while discarding other non-relevant words. However, the full sentences are split into halves without removing any word to prepare the non-causal pairs (X, Y) so that both X and Y contain linguistic flows. Fig. 2 illustrates a few lexical frequencies of the causal phrases in the training dataset. We notice from Fig. 2(a) that the most frequent five causal cue words are *if, because, cause, due to,* and *because of*. Fig. 2(b) and Fig. 2(c) display the top causal bigrams and effect bigrams, respectively. The causal bigrams are extracted from the causal phrases and similarly, the effect bigrams are extracted from the effect phrases. Intuitively, the frequent causal bigrams such as *risk uncertainty, new information* and *uncertainty factors* are part of causes in the causal relationships. Similarly, the frequent effect bigrams such as *actual results, differ materially* and *result differ* are the part of effects in the causal relationships. Fig. 3 and Fig. 4 illustrate the word co-occurrence heatmaps of the causal phrases and the effect phrases respectively, which also support the bigrams figures described above. We find that the dataset prepared by following the above approach is good enough to train a transfer learning model to achieve a comparable performance to the state-of-the-art approaches.

B. Pair-to-Sentence Conversion

After collecting the training dataset (refer to Section III-A) that contains both causal and non-causal pairs, we fine-tune a pretrained BERT model (pretrained on large dataset to capture general linguistic features) on this dataset. To do so, we first convert each causal pairs (X, Y) into a sentence by following the Hassanzadeh et al. [7]'s technique and represent the pair as "X may cause Y". For example, the pair *(Australian*



(a) Causal cue words

(b) Causal bigrams

(c) Effect bigrams

Fig. 2: Top 15 causal cue words, causal and effect bigrams

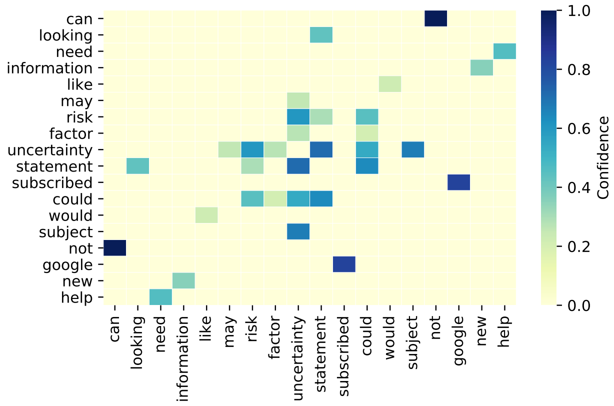


Fig. 3: Heatmap of causal words

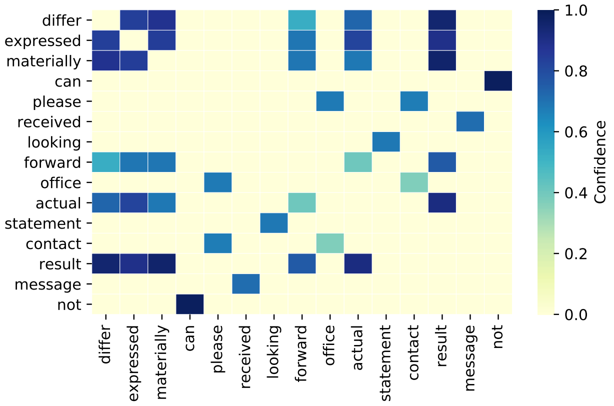


Fig. 4: Heatmap of effect words

fire, a jump in carbon concentrations in the atmosphere) is represented as "*Australian fire* may cause *a jump in carbon concentrations in the atmosphere*".

C. Fine-tuning BERT Model

BERT [15] is a powerful transformer-based language model. Recently, it has been used to build the state-of-the-art models for at least eleven natural language processing tasks. The BERT model is trained on a large dataset of BooKCorpus [16] (16GB) and English Wikipedia text. A pretrained BERT can capture the bidirectional representation of a sentence as it applies a deep bidirectional encoder mechanism. The power of BERT relies on the captured language representation during the pretraining phase. Though BERT model is trained for generic tasks such as context word and next sequence prediction, the model can also be fine-tuned to transfer its rich linguistic knowledge for more task specific needs. We believe that the rich linguistic knowledge patterns captured by BERT can be utilized to detect causality too. Hence, we apply a pretrained BERT model to answer binary causal questions by fine-tuning the model on our training dataset.

Before we perform the fine-tuning of the BERT model, each of the training pairs are converted into a sequence such as "X may cause Y" as described in Section III-B. The sequence is then passed to the BERT tokenizer which preprocesses the sentence into a format that is compatible to the BERT model. More specifically, the BERT tokenizer tokenizes each sentence into wordpieces [12] and each wordpiece is encoded with their corresponding dictionary indexes. The word dictionary is also prebuilt during the pretraining phase. The encoded wordpiece sequences are then passed to the BERT model for fine-tuning. Once the fine-tuning is done, the model is then used on the test datasets for evaluation.

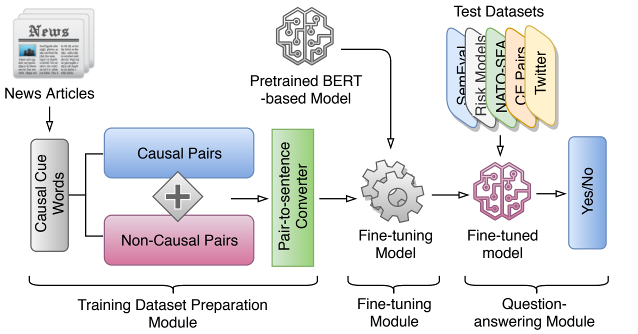


Fig. 5: Transfer learning based binary causal QA work flow

The overall workflow of our proposed BERT-based transfer learning approach for binary causal question answering framework is depicted in Fig. 5.

IV. Experiments

In this section, we describe our evaluation datasets and their source and attributes. We also discuss the benchmark approaches such as CausalNet approach [4] and NLM-BERT [7] that we have implemented to compare with our approach. The NLM-BERT model [7] was based on 17 million causal

TABLE III: Sample items from the SemEval dataset

Causal		Non_causal	
X	**Y**	**X**	**Y**
generator	energy	protein	researchers
lamp	light	rocks	pile
collision	fire	copper	tissue
attack	sorrow	article	criticisms
unemployment	alcoholism	drum	ear
gun	beam	work	difficulties

TABLE IV: Sample items from the NATO-SFA dataset

X	Y	Label
currency appreciation against US dollar	lower demand for USD	causal
country in state of war	fair justice system	non_causal
Technological dependencies	Vulnerabilities	causal
Fair burden sharing	increased cost of doing business	causal
weakening economic environment	foreign data providers dominate market	non_causal
Climate Change	New opportunities	causal

sentences and the two thresholds of the model are test dataset dependent as we discuss in Section II. In this paper, we improve the NLM-BERT model by proposing an approach for threshold selection that is not test dataset dependent, but can be learned from the training dataset. We have also implemented CausalNet approach [4] based on our own training dataset.

A. Evaluation Datasets

We evaluate our approach on five benchmark datasets - four datasets published by [7] and a Twitter dataset from our previous work [9]. Each dataset consists of pairs of texts such as (X, Y) and the corresponding labels. A brief description of these datasets is given below:

- **SemEval** - The SemEval dataset is a subset of the SemEval 2010 task 8 [17] dataset. The dataset is about identifying relationship between words in word pairs. Our SemEval dataset contains 1730 pairs which includes 865 *causal* and 865 *non_causal* pairs. The *causal* pairs are taken directly from the original dataset. The *non_causal* pairs are prepared by collecting pairs of words that has no causal relationship between them. Table III displays a few samples from the dataset. Any pairs in the table can be converted into a binary causal question. For instance, *(collision, fire)* is an example pair and “Could *collision* cause *fire*?” is the corresponding binary causal question.
- **NATO-SFA** - The NATO-SFA dataset is prepared from the Strategic Foresight Analysis (SFA) 2017 report [18] published by NATO (the North Atlantic Treaty Organization). The report includes a set of “trend” of changes in the world and their “implications” curated by human experts. The trends are considered as causes and the corresponding implications are considered as the effects. In the NATO-SFA dataset we have 118 pairs of words or phrases in total. There are 59 *causal* pairs and 59 *non_causal* pairs. The *non_causal* pairs are prepared by combining two phrases or words that has no direct cause-effect relationship in the report. Table IV shows a set of example pairs for the NATO-SFA dataset.

TABLE V: Sample items from the Risk Models dataset

X	Y	Label
Rising regional tensions	resource competition	non_causal
decrease in taxes	market disruptors	causal
growing social tension	reduced tourism	causal
increase in taxes	liberal politician or party elected	causal
strengthening economic environment	increased cost of doing business	non_causal
Natural disasters	Increased requirement for humanitarian support	causal

TABLE VI: Sample items from the CE pairs dataset

X	Y	Label
broadband access	more new businesses	causal
increased growth	dent consumer and business confidence	non_causal
consistent branding and pricing	increases revenue	causal
increase in corruption	negative effect on foreign investment	causal
management incentives	low investment	causal
increased government spending	decreased political stability	non_causal

- **Risk Models** - The Risk Models dataset is built using a set of models that is a part of a decision support system [19], [20]. Each model is a graph where the nodes represent an event and the edges represent a cause-effect relationship between two nodes. The dataset contains 402 *causal* pairs which are prepared from the relationships between edges and nodes. The dataset also has the same number of *non_causal* pairs that are prepared by combining unrelated edges and nodes. Table V illustrates a few sample pairs from the Risk Models dataset.
- **CE Pairs** - The CE (cause-effect) pairs dataset is an extension of the Risk Models dataset. To build the CE pairs dataset a set of node labels is assigned to 7 human annotators and asked to find corresponding cause or effect phrases by web search. The node labels of the model graphs are used as the cause or effect phrases and the corresponding cause of effect phrases are searched on the web. The dataset contains 302 pairs, where 50% of the pairs are *causal* and the remaining pairs are *non_causal*. Table VI shows a few example pairs from the dataset.
- **Twitter** - This dataset is prepared by Kayesh et al. [9] that contains labeled tweets on causality. The tweets in the dataset are related to Commonwealth Games, Gold Coast 2018. The dataset includes total 916 pairs with manually annotated labels. Table VII includes a set of example *causal* and *non_causal* pairs from this dataset.

B. Benchmark Approaches

We compare our approach to answer binary causal questions with a number of existing benchmark approaches. A brief description of the benchmark approaches is given below.

TABLE VII: Sample items from the Twitter dataset

X	Y	Label
i ned to be front and centre	it’s al about me	non_causal
families truly suport girl-child	we can se that sky to is not the limit	causal
you’re loking for us in the vilage	you’l know where to find us	non_causal
#comonwealthgames2018 and what did they do	people were urged to stay out of gold coast	non_causal
they are the best	both reached this point	causal
a mechanical isue	34am central to benleigh train has ben canceled	causal

1) NLM-BERT Model and its Variants: This section presents the NLM-BERT Model proposed by Hassanzadeh et al. [7] and its variants that we have implemented in our paper.

- NLM-BERT-17M [7] - This is the BERT-based natural language model proposed by Hassanzadeh et al. [7]. The approach uses 17M causal sentences as their training dataset. Since we couldn’t access their training dataset to reproduce the results we directly report the same thresholds and results mentioned in the original paper [7] on four test datasets: SemEval, NATO-SFA, Risk Models and CE Pairs. The similarity calculation of NLM-BERT and threshold selection are described later in this section.
- NLM-BERT [7] - To implement NLM-BERT that is local to our training dataset, we use the pretrained ‘Sentence-Transformer’ [21] model to transform each sentence into a 768-dimension vector. For indexing the sentence vectors and searching the closest k sentences, we use the faiss[4] library [22] as suggested by the authors [7]. For this model, we also use the same thresholds proposed by Hassanzadeh et al. [7] for all test datasets except the twitter dataset. For the twitter dataset, we have used the maximum of the thresholds proposed by Hassanzadeh et al. [7] for SemEval, NATO-SFA, Risk Models and CE Pairs.
- NLM-BERT++ - The implementation of this model is the same as the NLM-BERT model [7] except the usage of thresholds. Unlike the NLM-BERT model, we use a single pair of values for th_1 and th_2 which is 0.60 and 0.30, respectively, for all test datasets. These threshold values are determined automatically by following a technique described later in this section.

****NLM-BERT Similarity Scores Calculation****. The NLM-BERT model [7] is dependent on two similarity scores: *bert-sim-score* and *bert-c-score*. To calculate these scores, the model converts a candidate causal pair (X, Y) into two sentences “X may cause Y” and “Y may cause X”. Then, it calculates *bert-sim-score* and *bert-c-score* for “X may cause Y” and “Y may cause X”, respectively. To calculate *bert-sim-score*, the approach converts the sentence “X may cause Y” into a vector v_f using a pretrained BERT model and then, finds a set of k closest vectors $\{v_0, v_1, ..., v_{k-1}\}$ from the training dataset which has been converted into a list of vectors using the same pretrained BERT model. Then, it calculates *bert-sim-score* based on the average *cosine similarity* of the pairs (v_f, v_i), where $i \in \{0, 1, ..., k-1\}$ as given in Eq. 1.

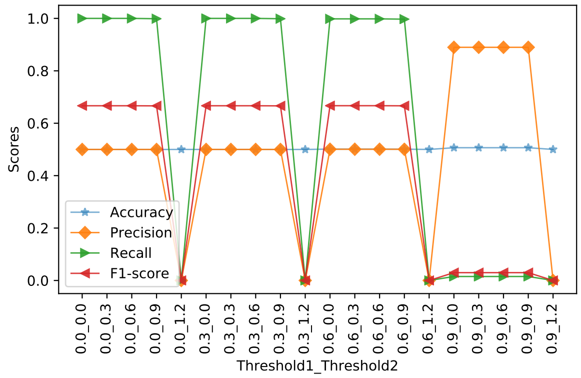


Fig. 6: Threshold selection for the NLM-BERT++ approach

$$\textit{bert-sim-score} = \frac{\sum_{i=0}^{k-1} CosineSimilarity(v_f, v_i)}{k} \tag{1}$$

To calculate *bert-c-score*, the reverse causal sentence “Y may cause X” is similarly converted into a vector v_r and the closest k sentence vectors $\{v_0, v_1, ..., v_{k-1}\}$ are extracted from the training dataset. The average *cosine similarity* score, which is labeled as *bert-reverse-sim-score*, is calculated based on the pairs (v_r, v_j), where $j \in \{0, 1, ..., k-1\}$ as given in Eq. 2. Finally, *bert-c-score* is calculated by dividing *bert-sim-score* by *bert-reverse-sim-score* as given in Eq. 3.

$$\textit{bert-reverse-sim-score} = \frac{\sum_{j=0}^{k-1} \text{CosineSimilarity}(v_r, v_j)}{k} \tag{2}$$

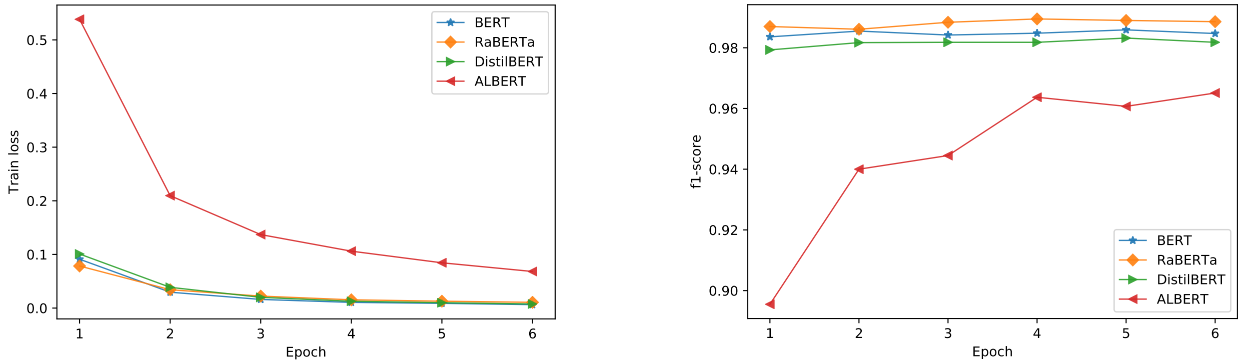
$$\textit{bert-c-score} = \frac{\textit{bert-sim-score}}{\textit{bert-reverse-sim-score}} \tag{3}$$

****Automatic Threshold Selection for NLM-BERT++****. In NLM-BERT model [7], the answer to the causal question “Could X cause Y?” depends on *bert-sim-score* and *bert-c-score*, which are maximized by thresholds th_1 and th_2, respectively. If *bert-sim-score* and *bert-c-score* is greater than th_1 and th_2, respectively, then the answer is *yes* and otherwise, the answer is *no* as given in Eq. 4.

$$f(X, Y) = \begin{cases} yes & \text{if } \textit{bert-sim-score} > th_1 \& \textit{bert-c-score} > th_2, \\ no & \text{otherwise} \end{cases} \tag{4}$$

In the original NLM-BERT approach [7], th_1 and th_2 need to be finalised for each test dataset, which requires expert supervision. For this, a prior knowledge of the test dataset characteristics is required. In this paper, we aim to eliminate this manual threshold selection process by automatically learning the thresholds from the training data (not the test dataset). We randomly split the training dataset into training and validation datasets. We use 75% data for training and the remaining 25% data for validation. Then, we follow the

[4] https://github.com/facebookresearch/faiss



(a) Training Loss

(b) Validation F1-score

Fig. 7: Pretraiend BERT-based models fine-tuning scores on the validation dataset

same procedure described in **NLM-BERT Scores Calculation** to calculate the two scores and use Eq. 4 to answer binary causal questions and evaluate them on the validation dataset for different combination of th_1 and th_2. For th_1, the value varies from 0.0 to 1.0 and the threshold th_2 ranges between 0.0 and 1.2. We set the step size to 0.3 that is used to increase these threshold values. Fig. 6 illustrates the validation scores for different combinations of thresholds. We observe that both f1-score and accuracy are maximized for $th_1 = 0.6$ and $th_2 = 0.3$ as it is evident from Fig. 6.

2) CausalNet Approach: The CausalNet approach [4] builds a causal network, which is a bidirectional graph where each node represents a word and each edge corresponds to a causal relationship, to capture the causal knowledge. Firstly, the training dataset is preprocessed using the set of causal cue words given in Table I. For a candidate causal pairs (X, Y), a causal score is calculated using the causal graph to decide whether X is a cause of Y. Firstly, a set of word pairs is prepared by implementing Cartesian product between words in both X and Y. For each pair, the necessity and sufficiency causal scores are calculated from the causal network, which are then combined together after maximizing by a threshold. Finally, the pair-wise scores are summed together and divided by the total number of words in X and Y to calculate the causal score. Interested readers are referred to [4] for detail explanation of the approach. In this experiment, we use the same 100K training causal pairs to build the causal network and considered a candidate pair (X, Y) as causal if the causal score is greater than zero.

3) Our Approach: This section presents the automatic binary question answering models that we have implemented based on the state-of-the-art BERT model and its variants by following the approach described in Section III.

- **BERT** [15] - BERT is a transformer-based language model proposed by Google Research. This model can be used for transfer learning based natural language processing (NLP) tasks. To implement this BERT model based transfer learning for our task, we use the transformers[5] library [23], which provides several pretrained transformers based language models. In our experiment, we use the 'bert-base-uncased' version of the pretrained model with 'BertForSequenceClassification' class.
- **RoBERTa** [24] - RoBERTa is a variation of the BERT model, which is proposed by Facebook AI. This approach focuses on hyperparameter tuning and alternative training strategies of BERT. The authors claim state-of-the-art performance on at least three publicly avaiable datsets. To implement this model in our experiment, we use the 'roberta-base' version of the pretrained model, which is released by the transformers library [23] with the 'RobertaForSequenceClassification' class.
- **DistilBERT** [25] - DistilBERT is a light-weight variation of BERT. The objective of this model is to reduce the number of hyper parameters and structural complexity of transformers while keeping the performance comparable to the original BERT model. The version of pretrained model we implement in this experiment is 'distilbert-base-uncased' and we use the "DistilBertForSequenceClassification" class to fine-tune the model.
- **ALBERT** [26] - ALBERT is another light-weight version of the BERT model. The ALBERT model is proposed to handle the issue of large training time, higher usage of memory and scaling for the large dataset. The approach applies parameter reduction technique to reduce run-time and improve memory usage. In this experiment, we fine-tune the 'albert-base-v2' version of the pretrained model with 'AlbertForSequenceClassification' class.

C. Experiment Settings

We perform the experiments in this paper on Google Colab GPU runtime, which offers 12GB of memory and a Tesla K80 GPU processor that has 2,496 CUDA cores. We run every BERT-based model by setting batch size to 32, maximum sequence length to 128, and learning rate to 0.00002. We use a small learning rate so that the pretrained weights are not overridden too much. Fig. 7 suggests that the training losses and validation f1-scores of BERT, RoBERTa and DistilBERT are plateau at around epoch 4 or 5. From these observations, we set the number of epochs to 5 for BERT, 4 for RoBERTa,

[5] https://github.com/huggingface/transformers

TABLE VIII: Pretrained BERT-based models fine-tuning scores

Model	Score	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5	Epoch 6
BERT	Train loss	0.0911	0.0296	0.0161	0.0106	0.0088	0.0064
	Val Acc	0.9842	0.986	0.9847	0.9855	0.9864	0.9853
	Val F1	0.9836	0.9855	0.9842	0.9848	**0.9859**	0.9847
RoBERTa	Train loss	0.0784	0.0342	0.0223	0.0154	0.0128	0.0107
	Val Acc	0.9876	0.9866	0.9888	0.9898	0.9893	0.989
	Val F1	0.987	0.9861	0.9884	**0.9895**	0.989	0.9886
DistilBERT	Train loss	0.1008	0.0387	0.02	0.0127	0.01	0.0076
	Val Acc	0.9799	0.9825	0.9823	0.9823	0.9839	0.9823
	Val F1	0.9793	0.9817	0.9818	0.9818	**0.9832**	0.9818
ALBERT	Train loss	0.5384	0.2096	0.137	0.1061	0.0844	0.0682
	Val Acc	0.8909	0.9396	0.9438	0.9647	0.9616	0.9658
	Val F1	0.8955	0.94	0.9445	0.9637	0.9607	**0.9651**

5 for DistilBERT, and 6 for ALBERT. Please refer to Table VIII to see the detail results of our validation experiments for choosing the number of epochs for each model.

D. Results and Discussion

In this section, we describe the experiment results for different experiment settings and test datasets. We compare our approach with different benchmark approaches and each of the benchmark approaches except NLM-BERT-17M are trained on the same training dataset that we describe in Section III-A.

Table IX displays the results of benchmark comparison. The results show that our transfer learning-based models outperform Hassanzadeh et al. [7]'s pretrained BERT and cosine similarity-based approach NLM-BERT model in terms of f1-score on at least 4 out of 5 test datasets. The NLM-BERT-17M model has the best f1-scores for NATO-SFA, Risk Models, and CE Pairs. However, when the model is trained on our smaller 100K causal pairs dataset with the original thresholds mentioned by the authors has zero true positive value on NATO-SFA and CE Pairs datasets. We get similar result on the Twitter dataset by using 0.94 and 0.90 as th_1 and th_2, respectively. NLM-BERT++ model outperforms the NLM-BERT model on all except the Risk Models dataset. In the Risk Models dataset, the f1-scores are comparable. Luo et al. [4]'s causal network based approach has varying results and its f1-scores and accuracy vary from 0.1590 to 0.6095 and from 0.49 to 0.5426, respectively. Comparing among the BERT-based models, BERT has the best f1-scores on SemEval and Risk Models datasets, DistilBERT has the best f1-scores on NATO-SFA and CE Pairs datasets, while RoBERTa achieves the best f1-score on the Twitter dataset.

From these experiment results we observe that Hassanzadesh et al. [7]'s NLM-BERT model is dependent on large training dataset and expert supervision for manual threshold selection. Also, the thresholds are not transferable as they are test dataset specific. In this paper we show that we can solve this manual threshold selection problem by automatically learning thresholds from training data. The automatically selected thresholds are not dependent to any test data hence they are transferable. We also observe that the transfer learning-based approaches can achieve comparable performance to NLM-BERT-17M when fine-tuned on a small training dataset. Although the training dataset is small the rich linguistic features captured by the pretrained transfer learning models enable them to achieve comparable performance. These results also demonstrate the effectiveness of our proposed training dataset preparation method where we automatically prepared a training dataset with 100K causal pairs and the same number of non-causal pairs without any human annotation process.

V. Conclusion

In this paper, we have proposed a transfer learning-based approach to answer binary causal questions. We have presented a semi-supervised training dataset preparation approach which automatically collects the training data from news articles using a set of causal cue words. We have shown how to fine-tune a pretrained BERT model on our training dataset to answer binary casual questions. Our approach achieves a comparable performance to a number of benchmark approaches on five benchmark test datasets that are extracted by human experts. From experiments, we observe that we can solve the issue of large training dataset requirement of many machine learning-based models by fine-tuning a pretrained transfer learning-based model on a carefully designed small training dataset.

References

[1] K. Radinsky, S. Davidovich, and S. Markovitch, "Learning causality for news events prediction," in *WWW*. ACM, 2012, pp. 909–918.

[2] M. Riaz and R. Girju, "Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations," in *SIGDIAL*, 2013, pp. 21–30.

[3] ——, "Recognizing causality in verb-noun pairs via noun and verb semantics," in *EACL*, 2014, pp. 48–57.

[4] Z. Luo, Y. Sha, K. Q. Zhu, S.-w. Hwang, and Z. Wang, "Commonsense Causal Reasoning between Short Texts," in *KR*, 2016, pp. 421–431.

[5] S. Sasaki, S. Takase, N. Inoue, N. Okazaki, and K. Inui, "Handling Multiword Expressions in Causality Estimation," *IWCS*, 2017.

[6] T. Dasgupta, R. Saha, L. Dey, and A. Naskar, "Automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks," in *SIGDIAL*, 2018, pp. 306–316.

[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*, 2019.

[8] R. Sharp, M. Surdeanu, P. Jansen, P. Clark, and M. Hammond, "Creating causal embeddings for question answering with minimal supervision," in *Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2016.

[9] H. Kayesh, M. Islam, and J. Wang, "On event causality detection in tweets," *arXiv preprint arXiv:1901.03526*, 2019.

TABLE IX: Comparison with other benchmark methods (NR: not reported, NA: not applicable)

Dataset	Method	# of Pairs	th1	th1	tp	fp	Acc	Pre	Rec	F1	AUC
SemEval	NLM-BERT-17M	17M	0.82	0.62	863	855	0.5050	0.5020	0.9980	0.6680	NR
	NLM-BERT	100K	0.82	0.62	225	63	**0.5936**	0.7812	0.2601	0.3903	**0.5936**
	NLM-BERT++		0.60	0.30	865	860	0.5029	0.5014	1.0000	0.6680	0.5029
	CausalNet		NA	NA	79	50	0.5168	0.6124	0.0913	0.1590	0.5168
	BERT		NA	NA	862	816	0.5266	0.5137	0.9965	**0.6779**	0.5266
	RoBERTa		NA	NA	865	864	0.5006	0.5003	1.0000	0.6669	0.5006
	DistilBERT		NA	NA	864	855	0.5052	0.5026	0.9988	0.6687	0.5052
	ALBERT		NA	NA	854	810	0.5254	0.5132	0.9873	0.6754	0.5254
NATO-SFA	NLM-BERT-17M	17M	0.94	0.82	58	53	**0.5420**	0.5230	0.9830	**0.6820**	NR
	NLM-BERT	100K	0.94	0.82	0	0	0.5000	0.0000	0.0000	0.0000	0.5000
	NLM-BERT++		0.60	0.30	59	59	0.5000	0.5000	1.0000	0.6667	0.5000
	CausalNet		NA	NA	20	16	0.5339	0.5556	0.3390	0.4211	**0.5339**
	BERT		NA	NA	55	55	0.5000	0.5000	0.9322	0.6509	0.5000
	RoBERTa		NA	NA	56	57	0.4915	0.4956	0.9492	0.6512	0.4915
	DistilBERT		NA	NA	57	57	0.5000	0.5000	0.9661	0.6590	0.5000
	ALBERT		NA	NA	53	57	0.4661	0.4818	0.8983	0.6272	0.4661
Risk Models	NLM-BERT-17M	17M	0.00	0.90	345	318	**0.5370**	0.5200	0.9380	**0.6690**	NR
	NLM-BERT	100K	0.00	0.90	402	402	0.5000	0.5000	1.0000	0.6667	0.5000
	NLM-BERT++		0.60	0.30	402	402	0.5000	0.5000	1.0000	0.6667	0.5000
	CausalNet		NA	NA	320	328	0.4900	0.4938	0.7960	0.6095	0.4900
	BERT		NA	NA	398	396	0.5025	0.5013	0.9900	0.6656	**0.5025**
	RoBERTa		NA	NA	390	392	0.4975	0.4987	0.9701	0.6588	0.4975
	DistilBERT		NA	NA	389	393	0.4950	0.4974	0.9677	0.6571	0.4950
	ALBERT		NA	NA	387	388	0.4988	0.4994	0.9627	0.6576	0.4988
CE Pairs	NLM-BERT-17M	17M	0.91	0.65	160	157	0.5090	0.5050	1.0000	**0.6710**	NR
	NLM-BERT	100K	0.91	0.65	0	0	0.5000	0.000	0.000	0.000	0.5000
	NLM-BERT++		0.60	0.30	160	160	0.5000	0.5000	1.0000	0.6667	0.5000
	CausalNet		NA	NA	99	93	**0.5188**	0.5156	0.6188	0.5625	**0.5187**
	BERT		NA	NA	155	153	0.5062	0.5032	0.9688	0.6624	0.5062
	RoBERTa		NA	NA	150	158	0.4750	0.4870	0.9375	0.6410	0.4750
	DistilBERT		NA	NA	158	154	0.5125	0.5064	0.9875	0.6695	0.5125
	ALBERT		NA	NA	147	155	0.4750	0.4868	0.9187	0.6364	0.4750
Twitter	NLM-BERT	100K	0.94	0.90	0	0	0.4989	0.000	0.000	0.000	0.5000
	NLM-BERT++		0.60	0.30	459	457	0.5011	0.5011	1.0000	0.6676	0.5000
	CausalNet		NA	NA	232	192	**0.5426**	0.5472	0.5054	0.5255	**0.5427**
	BERT		NA	NA	444	439	0.5044	0.5028	0.9673	0.6617	0.5034
	RoBERTa		NA	NA	451	438	0.5131	0.5073	0.9826	**0.6691**	0.5121
	DistilBERT		NA	NA	442	436	0.5055	0.5034	0.9630	0.6612	0.5045
	ALBERT		NA	NA	409	395	0.5142	0.5087	0.8911	0.6477	0.5134

[10] H. Kayesh, M. S. Islam, and J. Wang, “Event causality detection in tweets by context word extension and neural networks,” in *PDCAT*, 2019, pp. 352–357.

[11] ——, “A causality driven approach to adverse drug reactions detection in tweets,” in *ADMA*, 2019, pp. 316–330.

[12] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *CoRR*, vol. abs/1609.08144, 2016.

[13] R. Girju, “Automatic detection of causal relations for question answering,” in *ACL Workshop on Multilingual Summarization and Question Answering*, vol. 12, 2003, pp. 76–83.

[14] D. Corney, D. Albakour, M. Martinez, and S. Moussa, “What do a million news articles look like?” in *NewsIR*, 2016, pp. 42–47.

[15] 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.

[16] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” in *IEEE International Conference on Computer Vision*, 2015, pp. 19–27.

[17] 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 *Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, 2009, pp. 94–99.

[18] NATO, “Stratigic Foresight Analysis 2017 report,” https://www.act.nato.int/publications-ffao, 2017, [Online; accessed February 21, 2019].

[19] S. Sohrabi, A. V. Riabov, M. Katz, and O. Udrea, “An ai planning solution to scenario generation for enterprise risk management,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[20] S. Sohrabi, M. Katz, O. Hassanzadeh, O. Udrea, M. D. Feblowitz, and A. Riabov, “Ibm scenario planning advisor: Plan recognition as ai planning in practice,” *AI Communications*, no. Preprint, pp. 1–13, 2019.

[21] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.

[22] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with gpus,” *arXiv preprint arXiv:1702.08734*, 2017.

[23] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, “Huggingface’s transformers: State-of-the-art natural language processing,” *ArXiv*, vol. abs/1910.03771, 2019.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[25] 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.

[26] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” *arXiv preprint arXiv:1909.11942*, 2019.