


Article

Text to causal knowledge graph: A framework to synthesize knowledge from unstructured business texts into causal graphs

Seethalakshmi Gopalakrishnan¹, Victor Zitian Chen², Wenwen Dou¹, Gus Hahn-Powell³ , Sreekar Nedunuri⁴ and Wlodek Zadrozny⁵

¹ University of North Carolina at Charlotte; {sgopala4,wdou1}@uncc.edu

² Fidelity Investments; founder@gopeaks.org

³ University of Arizona; hahnpowell@arizona.edu

⁴ University of North Carolina at Charlotte; sreekarnedunuri@gmail.com

⁵ Computer Science Dept. and School of Data Science, UNC Charlotte, wzadrozny@uncc.edu

* Correspondence: sgopala4@uncc.com

Abstract: This article presents a state-of-the-art system to extract and synthesize causal statements from company reports into a directed causal graph. The extracted information is organized by the relevance to different stakeholder groups' benefits (customers, employees, investors, and the community/environment). The presented method of synthesizing extracted data into a knowledge graph comprises a framework that can be used for similar tasks in other domains, e.g. medical information. The current work addresses the problem of finding, organizing, and synthesizing a view of the cause-and-effect relationships based textual data, in order to inform and even prescribe the best actions that may affect target business outcomes related to different stakeholders' benefits (customers, employees, investors, and the community/environment).

Keywords: Causality extraction, Organizational data, Stakeholder Taxonomy, Natural Language Processing, NLP

1. Introduction

This article is motivated by the problem of extracting business knowledge from business texts, such as technical articles and financial reports.

The business problem is illustrated by the estimate by the International Federation of Accountants (IFAC), saying that the efforts of integrating various reporting data may have cost the financial industry alone \$780 billion annually [1]. Furthermore, in current climate, not only do managerial accountants and analysts need to monitor and analyze financial activities, but they also need to do so by making causal links to non-financial behaviors and outcomes such as sustainability, corporate social responsibility (CSR), environmental, social, and governance (ESG), and/or integrated reporting [2,3]. It is found by recent accounting research that decision-makers heavily rely on the causes-and-effects insights (e.g., materiality) underlying the accounting reporting to make financial and strategic decisions [4]. While more financial and non-financial reporting is making more information accessible about a company, it also creates an analytic paralysis for both internal decision-makers and external analysts to make sense of the complex and often hidden causal links among different KPIs (key performance indicators) and their drivers.

This article addresses the above business need by providing a framework, in which a collection of documents could be translated in to a knowledge graph representing causal relations expressed in the texts. Such a Text-to-Knowledge Graph tool will allow, for the timely synthesis of fragmented knowledge within and across reporting documents. Within enterprise performance management, this tool will assist human managers, auditors, accountants, and analysts to automatically detect, extract, and deconstruct causal propositions within and across company reports and then sort and visually connect the extracted causes and effects into a knowledge graph, which is a visual representation of variables ("entities") and their relationships ("links").

Contributions in this article

The article describes a prototype of the proposed framework, Text2CausalGraph, and evaluates its performance. Our specific contributions include the following:

- We created a new annotated dataset of causes-and-effect relationships and performance term classifications based on the S&P Financial Company 10-K reports.
- We created a pipeline to automatically read a text document and process it to create a knowledge graph.
- We compared the extracted causalities against a domain taxonomy and classify the extracted causalities.
- We have developed a novel approach to bridge machine reading with domain expertise (e.g., a pre-built taxonomy from domain experts)
- The presented architecture can be used as a framework for extracting causal information in other domains, for example in medical texts.

2. Related Work

Causality extraction is the process of extracting the cause and effect from a sentence. In the past few years, much work on causality extraction has been done, but still it remains a challenging task. A survey on the extraction of causal relations from text [5] categorizes the existing methodologies into *knowledge-based*, *statistical-machine-learning-based*, and *deep-learning-based methodologies*. We briefly show the diversity of these approaches below.

Earlier works in this area of causality extraction were using rules and linguistic features to extract cause/effect tuples [6], [7], [8]. Machine learning models can also be used to extract the causality from the text. Linguistic features, verb-pair rules, etc., as well as discourse features can be used to train the classifiers such as Naive Bayes and Support Vector machines [9,10]. In recent times deep learning-based models have been used for extracting the causalities from the text [11],[12],[13].

The causalities can be extracted in sentence level (intra-sentence) [14], [15], [16], [17], or it can also be extracted across the sentences (inter-sentence) [18], [19], [20]. A model can classify a sentence as causal based on the presence of an explicit connective (explicit causality) [21], [11], [13]. In the absence of causal connectives, semantic information can be used to find the causalities (which is called implicit causality) [22], [23].

A recent work on causality extraction [12] extends the SemEval 2010 Task 8 dataset by adding more data and uses BiLSTM-CRF with Flair embeddings [24] to extract cause/effect relationships. A similar work [25] uses CNN on the SemEval-2010 Task 8 dataset [26], Causal-TimeBank dataset [27], and Event StoryLine dataset [28], whereas [29] uses a Recursive Neural Tensor Network (RNTN) model [30]. Some of the works consider causality extraction as a span extraction or sequence labeling task [31]. CausalizeR [32] is a similar work that extracts the causal relationships from literature based on grammatical rules.

Finally, the emergence of large language models creates a new environment for extracting causality-related information. Some models such as GPT-3 may exhibit subpar performance (as shown in the Appendix B) to this article. On the other hand, GPT-4 has potential to outperform existing methods [33]. (At the time of this writing we do not have access to GPT4).

3. Data

We have collected and manually annotated the 2020 SEC 10-K Documents of 65 S&P Financial Companies. Five graduate students trained in business analytics, business administration, and/or economics were hired to manually annotate the causal insights from the documents based on a predefined dictionary of causal trigger words. At least two students carefully read each sentence with trigger words to ensure it describes a cause-and-effect relationship. Together, we have identified and manually annotated 2234 sentences that are causal in nature. For each of the identified causal sentences, the cause/effect relationship will be marked using the tags. Five graduate students manually annotated causes, triggers, and outcomes. After one round of discussions to resolve disagreements there was a 100%

inter-rater agreement.¹ An example sentence is given below:

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

Each of the identified causal relationships was mapped into two-level hierarchical taxonomy representing different stakeholder aspects of business performance indicators. Below is the stakeholder taxonomy we developed for representing business performance indicators.

Level 1	Level 2	Level 2 description
Performance (P)	Investors (INV)	The economic or financial outcome of the firm, which benefits investors, shareholders, debtholders, or financiers.
	Customers (CUS)	The value and utility of products/services the firm creates for and delivers to customers, clients, or users.
	Employees (EMP)	The benefits and welfare employees (workers and managers) receive from an organization.
	Society (SOC)	An organization's efforts and impacts on addressing community, environmental, and general public concerns.
	Unclassified	
Non-performance (NP)		Sentences which doesn't come under the performance category.

Table 1. Stakeholder Taxonomy which we use to classify the extracted causal statements. The causal statements extracted (Section 4.3) will be classified using the machine learning model (Section 4.4) into any one of the categories.

4. Methodology

The prototype, named Text2CausalGraph (meaning finding causal insights and converting them to a knowledge graph), comprises a series of machine learning modules to automatically detect, extract, label, and synthesize the causal insights from unstructured text in company reports.

The overall architecture is given in Figure 1. The system's operation includes four steps. Given a text document, the first step is to classify whether the given sentence is causal or not. From the list of the classified causal sentences, cause/effect will be extracted, and then the extracted causalities will be classified based on a stakeholder taxonomy. The final step is to visualize the classified taxonomy results (work in progress, and not reported in this article). In this pipeline, all the models are fine-tuned on manually annotated gold data, using transformer-based deep learning models.

¹ However, the 100% agreement does not imply complete consistency, e.g. some phrases included the determiner 'the' in some sentences but omitted them in others.

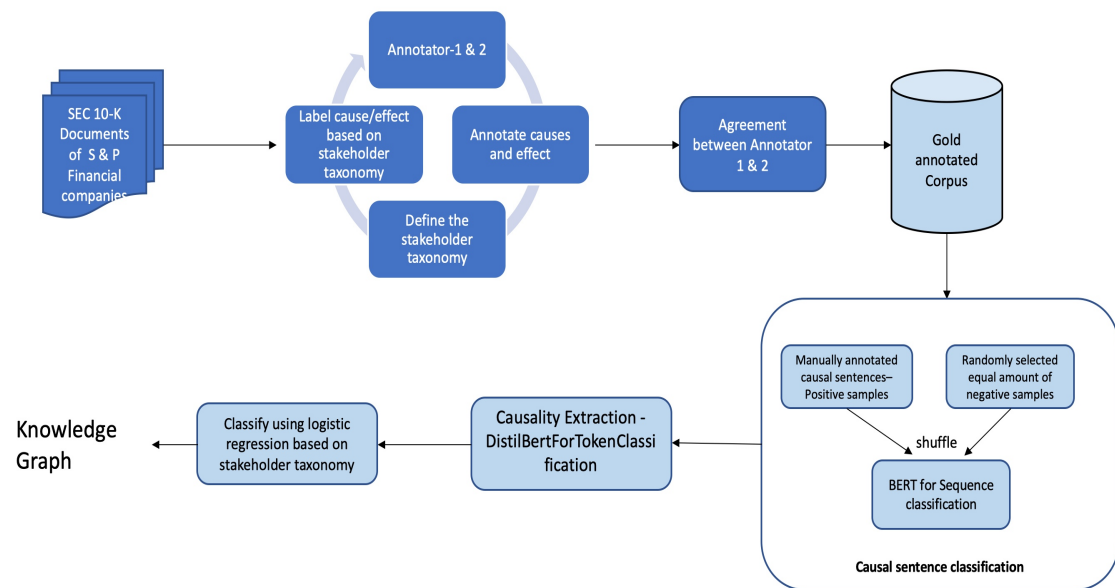


Figure 1. An architecture diagram showing the extraction and classification of causal statements from the S&P financial companies. The method consists of four main steps: causal sentence classification, causality extraction, classification of causalities based on stakeholder taxonomy, and construction of a knowledge graph. For this pipeline, the input is a set of text documents, and the output a knowledge graph showing the relations of cause and effect.

The following subsections explain the modules we have developed and the model's performance. The steps we follow in this process are given in Algorithm 1. The performance results, documented in Section 4.3, are based on splitting the manually-annotated data into training and testing portions.

Algorithm 1 Text to Knowledge Graph. The sample output of Algorithm 1 is shown in Fig. 2

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

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

Output: \mathcal{K}_G — A knowledge graph based on the stakeholder taxonomy from the causalities.

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

4.1. Data Preparation and Preprocessing

As an initial step, 62 reports for the year 2019 from the 10-K annual report documents of Standard & Poor (S&P 500) Financial Companies were retrieved from the Securities and Exchange Commission (SEC) website. On the SEC website, the data will be in Inline (eXtensible Business Reporting Language) XBRL format (IXBRL) format. We extracted this text in JSON format without filtering using Trafilatura [34], a Python package, and cleaned them using the NLTK package [35]. From this extracted and cleaned text, the causal

sentences are identified using the causal trigger words. The set of the causal trigger words we used for this task is given in Appendix A, which are adopted from [36]. If the trigger word is present in a sentence, it will be marked as a possible causal sentence using tags <causal-relation>. A JSON dataset with possible causal sentences marked was given to a set of graduate students to read the sentence and mark which part of the sentence is cause/effect/outcome. They were marked using the tags as given in Example 1.

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

4.2. Machine learning for automatic causal sentence detection and extraction

BERT [37] for sequence classification was fine-tuned on our dataset. As the training data, we labeled all the manually annotated causal sentences as causal. We randomly selected an equal amount of tweet data for the negative samples and shuffled both of them. Our data have an equal number of positive and negative samples. The obtained data was divided into train and test data. On the test data, the system obtained the F1 score of 87.65%.

4.3. Machine learning for automatic causality extraction

BERT is a state-of-the-art performing model for many NLP tasks, including Named Entity Recognition (NER). We used SpanBERT and DistilBERT models, adapted for token classification for causality extraction. Based on an 80% training set and 20% test set from the manually-annotated gold data, the performance of the SpanBERT model has the macro average F1-score of 0.89, the macro average precision of 0.87, and the macro average recall of 0.91 and DistilBERT has an average F1-score of 0.86, macro average precision of 0.81, and a macro average recall of 0.91.

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

Table 2. Summary of SpanBERT and DistilBERT's performance on the ORGanizational data for the Causality Extraction task (CE-ORG). Each token in the text is assigned a cause (C), effect (E), and Causal Trigger (CT) label. The results given above are obtained by splitting the manually annotated gold data into train and test partitions where the training partition is used to fine-tune BERT.

During the error analysis, we identified that in most places, stopwords are annotated as "O" and predicted as cause/effect or vice versa. In order to avoid this, as a post-processing step, we removed stop words from the list of tokens. We used the list of NLTK stopwords excluding negations. The results after removing the stop words are summarized in Table 3

	P(Span)	R(Span)	F1(Span)	P(Distil)	R(Distil)	F1(Distil)
Cause	0.83	0.88	0.85	0.79	0.87	0.83
Causal trigger	0.93	0.97	0.95	0.91	0.93	0.92
Effect	0.87	0.91	0.89	0.80	0.94	0.86

Table 3. Summary of the results of causality extraction after removing the stopwords from the list of the tokens. Here we use a set of common English stopwords that crucially omits negation tokens (ex. not) from the ignored set.

From the DistilBERT's performance after removing stopwords by comparing Table 2 and Table 3, we can understand that for the cause and effect, the F1-score is reduced if we remove the stop words, but for the causal triggers the F1-score increased from 0.81 to 0.92. SpanBERT's performance on the cause and effect slightly increases after removing the stopwords.

We also tried using the BERT-large model for the same causality extraction task. BERT-large got the macro average F1-score of 0.83, the macro precision of 0.78, and the macro recall of 0.90, which is lower than the DistilBERT and SpanBERT's performance.

Finally, we note that when using the BIO-label format to include the beginning and the inside tags for cause, effect, and trigger, we got a macro average F1-score of 0.60, an accuracy of 0.73, macro average precision of 0.73, and a macro average recall of 0.60 using DistilBERT. The results of running DistilBERT, SpanBERT, and BERT-large on our dataset are summarized in Table A1 in the Appendix.

4.4. Machine learning for automatic labeling of stakeholder taxonomy

We used the logistic regression model based on the performance in our prior experiments with similar data. Based on an 80% training set and 20% test set, with five-fold validation, the performance for the selected model has the average macro F1-score of 0.78, the accuracy of 0.89, the macro average precision of 0.76, and the average macro recall of 0.79 for Level 1. For Level 2, we got the macro average F1-score of 0.45, the accuracy of 0.88, the macro average precision of 0.47, and the macro average recall of 0.44.

	Precision	Recall	F1-Score	Support
Business Performance	0.58	0.65	0.62	12532
Business Non-performance	0.94	0.93	0.94	1976

Table 4. Performance of logistic regression model on Level 1 labels of the stakeholder taxonomy. The results summarized in this table are based on splitting the manually annotated data into train, test, and train a logistic regression model.

	Precision	Recall	F1-Score	Support
Customer	0.11	0.06	0.08	31
Employee	0.61	0.52	0.56	204
Investor	0.56	0.70	0.62	1013
Society	0.22	0.11	0.15	35
Unclassified	0.36	0.32	0.34	693
Business Non-performance	0.94	0.93	0.94	12532

Table 5. Performance of classification model on Level 2 labels of the stakeholder taxonomy. The results summarized in this table are based on splitting the manually annotated data into train, test, and train a machine learning model. In the manually annotated data, many of them are Non-performance labels, and we got a higher F1 score in that category.

From Table 1 and Table 5 we can understand that this is unbalanced data which is the reason for the poor performance of the logistic regression model on certain labels. (We are working on increasing the size of the dataset to have a balanced dataset). We also tried

running BERT. However, the results were low compared to using logistic regression: the average F1-score of 0.65 on Level 1 labels, and 0.28 on Level 2 labels.

4.5. Visualizing the output

The output of the process described in Section 4.4 is a table with the following information: a full sentence that is causal, cause/effect phrase, classification of cause/effect into Level1 and Level2 stakeholder taxonomy as given in Table 1. From this table, we can understand the relationships between the causes and the effects. This table can be converted into a directed graph or a knowledge graph. This is can be to visualized as the relationship between the causes, the effect and the labels in the taxonomy. A sample visualization for a sentence is given in Figure 2.

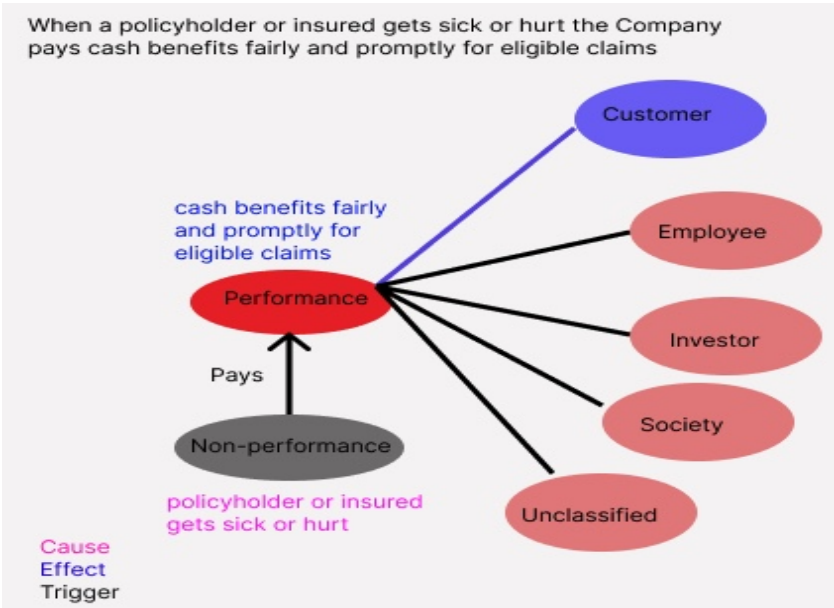


Figure 2. Knowledge graph for a sample sentence. Here the cause (policyholder or insured gets sick or hurt) is a non-performance label. The trigger word "pays" triggers an effect (cash benefits fairly and promptly for eligible claims). The effect is a performance label, and it can be further categorized into a customer at level 1 stakeholder taxonomy.

5. Error Analysis

We performed an error analysis on the DistilBERT model’s predictions on the test data. Below are some example errors from our predictions. In Example1, cause is predicted as effect; in Example 2, effect predicted as cause; and in Example 3, DistilBERT predicts CT wrongly and SpanBERT predicts CT correctly (an example of a pattern where DistilBERT performs poorly for causal triggers compared to SpanBERT).

Example 1.

Input text: *Over time, certain sectors of the financial services industry have become more concentrated as institutions involved in a broad range of financial services have been acquired by or merged into other firms. These developments could result in the Company’s competitors gaining greater capital and other resources, such as a broader range of products and services and geographic diversity. The Company may experience pricing pressures as a result of these factors and as some of its competitors seek to increase market share by reducing prices or paying higher rates of interest on deposits.*

Step 1 (extracting causal sentences using BERT) produces:

E E E E E E CT CT CT 0 0 C 0 0 E E C E E E E E E E E E E E E E E E E

Prediction - DistilBERT:

0 0 0 E 0 C E E E E E E E
Prediction - SpanBERT:
0 0 C T E E E E E E E

Overall,out of a total of 29674 tokens, 38% was E, 38% was of type C, 6% CT, and the rest of them were 0. Based on the error analysis results, 5% of E predicted as C, 1% of E predicted as 0, 11% of C predicted as E, 3% of C predicted as 0, 2% of CT predicted as E, 3% of CT predicted as 0. Less than 1% of E predicted as CT, C predicted as CT, CT predicted as C.

6. Conclusion

In this article we presented a framework for extracting a causal knowledge graph from text documents. Secondly we described a prototype, Text2Graph, applying this framework to organizational performance and financial data which we curated as part of the project. We also showed how to integrate extracted causalities into a a stakeholders taxonomy. The results show the feasibility of causal information extraction and the conversion of this information into a potentially actionable knowledge graph. This is the first step in addressing the needs of business analysts by integrating information from multiple textual sources into a single knowledge model.

Author Contributions: S.G performed the majority of the experiments and contributed to writing. V.C directed the data development, project development, and design of the pipeline. W.D oversaw the project activities and provided feedback throughout the development cycle. G.HP conceptualized the approach for causal relation extraction and stakeholder taxonomy classification. S.N Worked on data preparation, processing part and developed a causal sentence classification model; all his work was during his studies at UNCC. W.Z designed parts of experiments, provided feedback and contributed to writing.

Funding: This research was partly funded by National Science Foundation (NSF) grant number 2141124.

Data Availability Statement: Our data and implementation are available at <redacted; to be released upon publication>

Acknowledgments: The authors acknowledge the help of Yifei (Alice) Dong, Golnaz Dadgar, Ryan Hammond, Punit Mashruwala, Ravi Duvvuri in data preparation/annotation

Conflicts of Interest: Gus Hahn-Powell declares a financial interest in Lum AI. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies. Victor Z. Chen notes that the reported research was done independently from and does not represent the author’s affiliation. The other authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
IFAC	International Federation of Accountants
CSR	Corporate Social Responsibility
ESG	Environmental, Social, and Governance
SEM	Structural Equation Modeling
SCITE	Self-attentive BiLSTM-CRF wIth Transferred Embeddings
BiLSTM-CRF	Bidirectional Long Short-Term Memory-Conditional Random Field
CNN	Convolutional neural network
NLTK	Natural Language Toolkit
SEC	Securities and Exchange Commission
BERT	Bidirectional Encoder Representations from Transformers

Appendix A

Appendix A.1

The results of the causality extraction in Bio label format are summarized below. Here in most cases, the model cannot differentiate between the B-C/B-E and I-C/I-E.

	Precision	Recall	F1-Score
Beginning of effect	0.67	0.06	0.10
Beginning of cause	0.68	0.27	0.39
Inside of cause	0.76	0.83	0.79
Inside of causal trigger	0.76	0.95	0.84
Inside of effect	0.72	0.94	0.82
Beginning of causal trigger	0.89	0.85	0.87

Table A1. Summary of DistilBERT's performance on causality extraction task. Each token in the text is assigned a BIO label. The results given above are obtained by splitting the manually annotated gold data into train and test partitions where the training partition is used to fine-tune BERT.

Appendix A.2

	Precision	Recall	F1-Score
Beginning of effect	0.62	0.63	0.62
Beginning of cause	0.56	0.59	0.57
Inside of cause	0.78	0.87	0.83
Inside of causal trigger	0.94	0.96	0.95
Inside of effect	0.84	0.90	0.87
Beginning of causal trigger	0.94	0.96	0.95

Table A2. Summary of SpanBERT's performance on causality extraction task. Each token in the text is assigned a BIO label. The results given above are obtained by splitting the manually annotated gold data into train and test partitions where the training partition is used to fine-tune BERT.

Appendix A.3

	Precision	Recall	F1-Score
Beginning of effect	1.00	0.00	0.00
Beginning of cause	0.83	0.02	0.04
Inside of cause	0.70	0.87	0.77
Inside of causal trigger	0.63	0.70	0.66
Inside of effect	0.71	0.91	0.80
Beginning of causal trigger	0.74	0.67	0.70

Table A3. Summary of BERT-large performance on causality extraction task. Each token in the text is assigned a BIO label. The results given above are obtained by splitting the manually annotated gold data into train and test partitions where the training partition is used to fine-tune BERT.

Appendix B

In recent times, prompting large language models has given state-of-the-art performing results for many NLP tasks [38,39]. We tried a few-shot prompting of GPT3 on a sample of 100 sentences from our dataset. The model's results are summarized in Table A4. At the time of running these experiments, we don't have access to GPT4.

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

Table A4. Few-shot prompting of GPT3 on the organizational causality extraction dataset. This result is on the sample 100 sentences from the dataset.

	Precision	Recall	F1-Score
Non-Performance	0.72	0.80	0.76
Performance	0.12	0.08	0.10

Table A5. Few-shot prompting of GPT3 on the Level1 labels of the stakeholder taxonomy . This result is on the sample 100 sentences from the dataset.

References

1. IFAC.; International Federation of Accountants. Regulatory Divergence: Costs, Risks and Impacts. <https://www.ifac.org/knowledge-gateway/contributing-global-economy/publications/regulatory-divergence-costs-risks-and-impacts>, 2018.
2. Khan, M.; Serafeim, G.; Yoon, A. Corporate sustainability: First evidence on materiality. *The accounting review* **2016**, *91*, 1697–1724.
3. Naughton, J.P.; Wang, C.; Yeung, I. Investor sentiment for corporate social performance. *The Accounting Review* **2019**, *94*, 401–420.
4. Green, W.J.; Cheng, M.M. Materiality judgments in an integrated reporting setting: The effect of strategic relevance and strategy map. *Accounting, Organizations and Society* **2019**, *73*, 1–14.
5. Yang, J.; Han, S.C.; Poon, J. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems* **2022**, pp. 1–26.
6. Radinsky, K.; Davidovich, S.; Markovitch, S. Learning causality for news events prediction. In Proceedings of the Proceedings of the 21st International Conference on World Wide Web, 2012, pp. 909–918.
7. Ittoo, A.; Bouma, G. Minimally-supervised learning of domain-specific causal relations using an open-domain corpus as knowledge base. *Data & Knowledge Engineering* **2013**, *88*, 142–163.
8. Kang, N.; Singh, B.; Bui, C.; Afzal, Z.; van Mulligen, E.M.; Kors, J.A. Knowledge-based extraction of adverse drug events from biomedical text. *BMC Bioinformatics* **2014**, *15*, 1–8.
9. Pechsiri, C.; Kawtrakul, A.; Piriyaikul, R. Mining Causality Knowledge from Textual Data. In Proceedings of the Artificial Intelligence and Applications, 2006, pp. 85–90.
10. Keskes, I.; Zitoune, F.B.; Belguith, L.H. Learning explicit and implicit arabic discourse relations. *Journal of King Saud University-Computer and Information Sciences* **2014**, *26*, 398–416.
11. Xu, Y.; Mou, L.; Li, G.; Chen, Y.; Peng, H.; Jin, Z. Classifying relations via long short term memory networks along shortest dependency paths. In Proceedings of the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1785–1794.
12. Li, Z.; Li, Q.; Zou, X.; Ren, J. Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings. *Neurocomputing* **2021**, *423*, 207–219.
13. Wang, L.; Cao, Z.; De Melo, G.; Liu, Z. Relation classification via multi-level attention cnns. In Proceedings of the Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1298–1307.
14. Garcia, D.; EDF-DER.; IMA-TIEM. COATIS, an NLP system to locate expressions of actions connected by causality links. In Proceedings of the Knowledge Acquisition, Modeling and Management: 10th European Workshop, EKAW'97 Sant Feliu de Guixols, Catalonia, Spain October 15–18, 1997 Proceedings 10. Springer, 1997, pp. 347–352.
15. Khoo, C.S.; Chan, S.; Niu, Y. Extracting causal knowledge from a medical database using graphical patterns. In Proceedings of the Proceedings of the 38th annual meeting of the association for computational linguistics, 2000, pp. 336–343.
16. Pakray, P.; Gelbukh, A. An open-domain cause-effect relation detection from paired nominals. In Proceedings of the Nature-Inspired Computation and Machine Learning: 13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16–22, 2014. Proceedings, Part II 13. Springer, 2014, pp. 263–271.
17. Smirnova, A.; Cudré-Mauroux, P. Relation extraction using distant supervision: A survey. *ACM Computing Surveys (CSUR)* **2018**, *51*, 1–35.
18. Marcu, D.; Echihiabi, A. An unsupervised approach to recognizing discourse relations. In Proceedings of the Proceedings of the 40th annual meeting of the association for computational linguistics, 2002, pp. 368–375.
19. Jin, X.; Wang, X.; Luo, X.; Huang, S.; Gu, S. Inter-sentence and implicit causality extraction from chinese corpus. In Proceedings of the Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I 24. Springer, 2020, pp. 739–751.

20. Oh, J.H.; Torisawa, K.; Hashimoto, C.; Sano, M.; De Saeger, S.; Ohtake, K. Why-question answering using intra-and inter-sentential causal relations. In Proceedings of the Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013, pp. 1733–1743.
21. Girju, R. Automatic detection of causal relations for question answering. In Proceedings of the Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering, 2003, pp. 76–83.
22. Martínez-Cámara, E.; Shwartz, V.; Gurevych, I.; Dagan, I. Neural disambiguation of causal lexical markers based on context. In Proceedings of the IWCS 2017—12th International Conference on Computational Semantics—Short papers, 2017.
23. Ittoo, A.; Bouma, G. Extracting explicit and implicit causal relations from sparse, domain-specific texts. In Proceedings of the Natural Language Processing and Information Systems: 16th International Conference on Applications of Natural Language to Information Systems, NLDB 2011, Alicante, Spain, June 28-30, 2011. Proceedings 16. Springer, 2011, pp. 52–63.
24. Akbik, A.; Bergmann, T.; Blythe, D.; Rasul, K.; Schweter, S.; Vollgraf, R. FLAIR: An easy-to-use framework for state-of-the-art NLP. In Proceedings of the Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations), 2019, pp. 54–59.
25. Li, P.; Mao, K. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications* **2019**, *115*, 512–523.
26. Hendrickx, I.; Kim, S.N.; Kozareva, Z.; Nakov, P.; Séaghdha, D.O.; Padó, S.; Pennacchiotti, M.; Romano, L.; Szpakowicz, S. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *arXiv preprint arXiv:1911.10422* **2019**.
27. Mirza, P. Extracting temporal and causal relations between events. In Proceedings of the Proceedings of the ACL 2014 Student Research Workshop, 2014, pp. 10–17.
28. Caselli, T.; Vossen, P. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In Proceedings of the Proceedings of the Events and Stories in the News Workshop, 2017, pp. 77–86.
29. Fischbach, J.; Springer, T.; Frattini, J.; Femmer, H.; Vogelsang, A.; Mendez, D. Fine-grained causality extraction from natural language requirements using recursive neural tensor networks. In Proceedings of the 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW). IEEE, 2021, pp. 60–69.
30. Socher, R.; Lin, C.C.; Manning, C.; Ng, A.Y. Parsing natural scenes and natural language with recursive neural networks. In Proceedings of the Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 129–136.
31. Lyu, C.; Ji, T.; Sun, Q.; Zhou, L. DCU-Lorcan at FinCausal 2022: Span-based Causality Extraction from Financial Documents using Pre-trained Language Models. In Proceedings of the Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022, 2022, pp. 116–120.
32. Ancin-Murguzur, F.J.; Hausner, V.H. causalizeR: a text mining algorithm to identify causal relationships in scientific literature. *PeerJ* **2021**, *9*, e11850.
33. Kicman, E.; Ness, R.; Sharma, A.; Tan, C. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. *arXiv preprint arXiv:2305.00050* **2023**.
34. Barbaresi, A. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, 2021, pp. 122–131.
35. Bird, S. NLTK: the natural language toolkit. In Proceedings of the Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, 2006, pp. 69–72.
36. Barrett, E.; Paradis, J.; Perelman, L.C. The Mayfield Handbook of Technical & Scientific Writing. *Mountain View, CA: Mayfield Company* **1998**.
37. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.
38. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large Language Models Encode Clinical Knowledge. *arXiv preprint arXiv:2212.13138* **2022**.
39. Si, C.; Gan, Z.; Yang, Z.; Wang, S.; Wang, J.; Boyd-Graber, J.; Wang, L. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150* **2022**.