PolitiCause: An Annotation Scheme and Corpus for Causality in Political Texts

Paulina Garcia-Corral, Hannah Béchara, Ran Zhang, Slava Jankin
Data Science Lab - Hertie School
Natural Language Learning Group - University of Mannheim
School of Government - University of Birmingham
corral@hertie-school.org, bechara@hertie-school.org, ran.zhang@uni-mannheim.de, v.jankin@bham.ac.uk

Abstract

In this paper, we present PolitiCAUSE, a new corpus of political texts annotated for causality. We provide a detailed and robust annotation scheme for annotating two types of information: (1) whether a sentence contains a causal relation or not, and (2) the spans of text that correspond to the cause and effect components of the causal relation. We also provide statistics and analysis of the corpus, and outline the difficulties and limitations of the task. Finally, we test out three transformer-based classification models on our dataset as a form of evaluation. The models achieve a moderate performance on the dataset, with a MCC score of 0.62. Our results show that PolitiCAUSE is a valuable resource for studying causality in texts, especially in the domain of political discourse, and that there is still room for improvement in developing more accurate and robust methods for this problem.

Keywords: causal text mining, causal argument, political text

1. Introduction

Causal language refers to the use of expressions that convey causal relationships in text. This language can be complex and can span across multiple sentences. It can be expressed using discourse connectives such as "because", but can also be conveyed via causal verbs. Furthermore, causal language can be implicit and understood via contextual interpretation (Solstad and Bott, 2017). Additionally, events can be causally related in text without real world causation, and expressed as hypothetical or counterfactual. Causal reasoning is a fundamental aspect of human cognition that is linked to action and intervention (Sloman, 2005). It is essential for many cognitive and social tasks, such as decision making, planning, explanation, prediction, and argumentation.

The automatic extraction of causal arguments is a high-value task, as it enables the extraction of relationships that can be utilized for downstream applications. Causal relation extraction can be employed to model causal information, by creating causal chains or causal networks, in graph form. These can then be used for news understanding, text summarization, question-answering, and common sense reasoning (Drury et al., 2022). Additionally, the information extracted from causal relations in text can be used to make predictions, for example, in early warning systems or disaster management.

Much of the previous research into causal language detection is limited to scientific language. Examples include differentiation between causal and correlational language in scientific publications, and identifying symptoms and side effects in medical trials. However, causal relationships established in less scientific settings are expressed using different linguistic and syntactical structures.

In this paper, we introduce PolitiCAUSE[1], a causal language corpus focused on causal structures in political language. In political text, identifying causal relationships is crucial to analyze policy argumentation and fact-check political communication (Vössing, 2023; Falk and Lapesa, 2022; Reisert et al., 2018). The expressions used in this context are entrenched in a rich tradition of political rhetoric that is known for its persuasive objective. For causal language, this means arguing about potential effects of policy interventions, as well as counterfactual statements of previous policy decisions. Current annotation schemes are not tailored to cover these causal language expressions typical of political argumentation. Therefore, we develop an annotation scheme that specifically targets causal constructions in political texts and use it to annotate sentences collected from two different political corpora: The United Nations General Debate Corpus (UNGDC), and press conference transcriptions from the United Kingdom (UKPress).

The extraction of causal language relations has evolved significantly from the early days when linguistic pattern recognition was the primary approach. The advent of deep learning, particularly transformer-based models, has enabled supervised learning methods to improve results. However, the primary limitations persist, including the difficulty in extracting relations that span across sentences and the implicit or incomplete causal relations in sentences. Moreover, these models need high quality human-annotated datasets, usually costly and time consuming to produce.

For causal relation extraction, the standard approach is two-fold (Yang et al., 2022; Drury et al., 2022):

1. causal classification: identifying units of text that have causal language. This can be at the word,

[1] https://github.com/pgarco/PolitiCAUSE

LREC-COLING 2024, pages 12836–12845

sentence or paragraph level.

2. cause-effect span detection: tagging spans of text as “cause” and “effect”. We refer to this as “causal tagging”.

To implement this two-fold approach, data must undergo two stages of annotation: first, the text unit must be labeled as causal or not causal, and second, text units labeled as “causal” must be tagged with cause and effect. The option of establishing a relationship tag between spans is also available.
The rest of this paper is organized as follows: In Section 2, we provide an overview of previous research into causal detection in the form of previously curated causal datasets. In Section 3 we detail the design of our annotation scheme, its guidelines, and the annotation rules. In Section 4, we describe how we collected the sentences that make up our PolitiCAUSE dataset, detail the annotation process, provide some statistics about the annotated corpus, and we evaluate the corpus by benchmarking the data using transformer-based models for sequence classification, also providing an error analysis. Finally, we sum up our paper in Section 5.

2. Related work

Over the past decade, causal text mining has emerged as an increasingly important task, distinct from the general argument mining field. As a result, significant efforts have been made to create and test datasets annotated for causality. Based on their task, existing datasets can be classified into three categories: (1) Datasets for causal text mining, (2) Datasets for non-computational linguistic applications, and (3) Question-Answering datasets for causal inference language models.

2.1. Datasets for Causal Text Mining

- The BeCAUSE 2.0 (Dunietz et al., 2017) corpus is a dataset of causal language that contains 5,380 samples, annotated based on construction grammar. It contains annotation spans for “Cause”, “Effect” and “Connective” in single sentences. It includes newspaper articles, a random selection of the Penn Discourse Treebank (Prasad et al., 2007), and transcriptions of the US Congress. The main disadvantages is that it has complex annotation rules that yield low inter-annotator agreement for non-experts, and it’s a small dataset.

- The Parallel Wikipedia Corpus (Hidey and McKeown, 2016) is a collection of text samples that have been annotated for causal connectives with alternative lexicalizations. The authors identified common causal markers and searched for pairs between parallel English and Simple Wikipedia articles, resulting in 265,627 causal connective pairs. However, the main limitation of this corpus is that it does not take into account signals in causal relations and does not limit the size of cause and effect spans between the causal connectives.

- Causal-TimeBank (Mirza et al., 2014) is a dataset from the TempEval-3 task (UzZaman et al., 2014) that has been annotated with causal signals and causal links, in addition to the temporality annotations it already contains. However, the approach used to create this dataset yielded low precision mainly due to the presence of non-causal connectors in the data.

- The EventStoryLine Corpus (Caselli and Vossen, 2017) contains 258 documents annotated for both temporal and causal language for the extraction and classification of events in stories. It includes annotations for both explicit and implicit causal relations. It’s main limitation is that it only includes 117 explicit causal relations.

- The Causal News Corpus (Tan et al., 2022) was created to include both explicit relations and clause-based arguments. It comprises 3,559 event sentences from protest event news. Each sentence is labeled to indicate whether it contains causal relations or not. The authors achieved an F1 score of 83.46% in 5-fold cross-validation using a transformer-based model. Furthermore, the corpus is transferable to Causal-TimeBank and the Penn Discourse Treebank.

- The Penn Discourse Treebank (PDTB) (Prasad et al., 2019), SemEval-2010 Task 8 (Hendrickx et al., 2010) and SemEval-2007 Task 4 (Girju et al., 2007), are all large datasets annotated for multiple argument types, including some causal language markers. The PDTB includes 9,190 causal examples from the Wall Street Journal articles. SemEval-2010 Task 8 has an binary-labeled dataset for 10 types of relations, 12.4% of which are cause-effect. SemEval-2007 Task 4 includes 140 training examples with 52% positive data and 80 test examples with 51% positive data for Cause-Effect relations.

- UniCausal recently introduced by Tan et al. (2023) is a dataset for causal text mining that unifies six high-quality datasets: Parallel Wikipedia, BeCAUSE, Causal-TimeBank, EventStoryLine, PDTB, and SemEval 2010 Task 8. The resulting dataset contains 58,720 sentences for causal identification, 12,144 for cause-effect span detection, and 69,165 examples for causal pair classification. This effort is the largest to date for causal text mining.

2.2. Datasets for Non-Computational Linguistic Applications

- Gu et al. (2016) introduced Chemical Induced Disease (CID) relations extraction corpus, which contains relations between drugs and their adverse effects. The authors use various linguistic

features to train maximum entropy models for relation classification. The system achieves an F-scores of 60.4% and 58.3% on the development and test datasets, respectively.

- Mariko et al. (2021) compiled FinCausal, a dataset of financial documents annotated for causal relations. The paper presents the results of 16 participating teams, and discusses the challenges of causality detection in the financial domain including the complexity and diversity of causal expressions, the ambiguity and inconsistency of causal annotations, and the scarcity and imbalance of annotated data for causal extraction.

- Yu et al. (2019) proposed a system to differentiate between correlational and causal language in scientific publications. The authors attempt to address the problem of inappropriate causal interpretation of correlational findings from observational studies. They develop a BERT-based prediction model trained on an annotated corpus of over 3,000 PubMed research conclusion sentences. They report an accuracy of 0.90 and a macro-F1 score of 0.88 on the annotated corpus.

2.3. Datasets for Causal Inference

- Du et al. (2022) created e-CARE, a causal language dataset for question-answering and inference tasks. However, it was not developed for text mining, but rather for real world causation: it aims to answer *what* is the cause of something, rather than to detect whether there is a causal relationship between events in a text (which may be false in reality). Nevertheless, this data can be adapted for causal relation extraction tasks with some modifications.

3. PolitiCAUSE's Annotation scheme

The main objective of our annotation scheme is to identify explicit causal relations in single sentences. Our annotation scheme is not designed to capture implicit causality, incomplete causal structures or intersentential relationships in text. We base our annotation scheme on the view of causality as a psychological concept that is imperfectly expressed in language (Neeleman and Van de Koot, 2012). The complexity of causal language poses challenges for strict pattern-based annotation approaches. Therefore, our codebook's main objective is to enable coders to interpret sequences without relying on pattern recognition or grammatical rules to guide their span selection; instead, the training's main goal is to develop a shared understanding of what causal relations are.

3.1. Annotation Rules

We define causal relations in text as the explicit mention of a relationship where one event causes another event to happen. We introduce some causal discourse connectors, such as "because" or "therefore" as examples of signal or trigger words that can facilitate the identification of causal relations. However, we do not require the annotators to learn or annotate them. The annotation is based on the interpretation of the text.
The definition is followed by a first simple example as an introduction to the task:

I did not eat because I forgot my lunch.

Example 1: A sentence with two event, Event 1 in green is tagged as the "cause" and Event 2, in yellow, is tagged as the "effect".

The connector *because* signals a causal relationship between Event 1 and Event 2[2]. Forgetting lunch is the cause of not eating, the text expresses that there is a causal link between Event 1 and its consequence, Event 2.
Furthermore, we asked annotators to rephrase the sentence into a *Because-first* structure, and evaluate if extra information needed to be added for the sentence to be complete as a test to identify explicit causal relationships in sentences. We demonstrate this in the following example:

Peru lost many pre-Hispanic treasures as a result of Spanish colonial looting.

Because of Spanish colonial looting, Peru lost many pre-Hispanic treasures.

Example 2: The sentence is rephrased to a *Because-first* structure to test if it contains a complete causal relationship. Event 1 in green is tagged as the "cause" and Event 2, in yellow, is tagged as the "effect". No extra information is needed to rephrase, hence there is a complete causal structure.

Formally, the annotation scheme consisted of 5 steps:

1. Identify if there is a causal relationship connecting an Event 1 to a change of state in another Event 2. If this connection is present, label as causal. Otherwise, label as not causal.

2. If a causal relationship is found, tag the relevant text spans as "cause" and "effect".

3. Establish a relationship token between the "cause" and "effect" spans.

4. If there is a subject to the cause or effect, select the text span and tag as "subject", then, estab-

[2]This paper follows green for "cause", and yellow for "effect" as the color scheme for causal tagging.

lish a "subject" relationship token between subject and cause or effect.[3]

5. Rate your confidence score on a scale from 1 to 5.

Causal relations also included potential causes and effects, as well as counterfactual statements for the positive class (See Figure 3.1). Political communication is used to argue and convince, and relies on presenting future scenarios of policy impact or blame attribution of current issues to past policy, both of which are important to capture for downstream applications. Hence, we asked the annotators to search for sentences that establish causal relations, even if they have not happened yet or can't be proven to happen in the real world.

A. The Embargo Act did not improve America's diplomatic position.

B. Our country needs to come together to overcome the COVID-19 crisis.

Example 3: Political communication often includes arguments about failed policies, in sentence A there is a causal relationship between "the Embargo Act" and "not improve"; It also uses arguments to win policy support, in sentence B there is a causal relation between "come together" and "overcome the COVID-19 crisis".

3.2. Extended guidelines

We instructed the annotators not to use any contextual information during annotation, and to establish causal relationships even if the assertions was factually untrue. If a causal relation was stated in the text, it was considered to exist textually, regardless of real world causality. A causal relation had to be complete (a cause and an effect event had to be inside of the same sentence) and explicit (can not use outside information to complete the sense of the relationship) in the text to label a full sequence as "causal".

In addition to the causal relation annotation, Step 4 of the annotation scheme required the annotators to add a "subject" span if present, which is not the grammatical subject. The main purpose of this step was to reduce the disagreement caused regarding what to include as part of the "cause" or "effect", since this a point of major inter-annotator disagreement in other causal language datasets. In our annotation scheme "subject" only answer the question "*what entity* causes the event?" and "*what entity* is affected?".

[3]We refer to "subject" as any causer entity or affected entity, this is not the grammatical subject.

	tokens (word-level)		
	N	mean	std
UNGD	8,872	2,702.87	1,357.05
UK Press	429	787.78	477.58

Table 1: Description of the two sampled corpora. Where N is the number of documents in each corpus, and the token statistics are per document.

Step 5 was also included to measure the annotators' confidence level for each sentence and filter out sentences that had a low mean confidence score. Political rhetoric can be misleading by design to allow space for speculation, and we consider that annotators are not exempt from intentional ambiguity.

4. PolitiCAUSE Corpus

4.1. Building the PolitiCAUSE Corpus

We constructed a corpus of political texts by sampling from two political corpora: The United Nations General Debate Corpus (UNGD) (Jankin et al., 2023) and the United Kingdom Press Conference Transcriptions (UKPress). The UNGD Corpus comprises statements delivered by heads of state or government of all UN member states at the annual UN General Assembly sessions. The UNGD contains official English translations of the multilingual UN General Debate statements, provided by the UN Library. To complement the scripted nature of the UNGD, we included the UKPress data, which contains more conversational and lively political language. We obtained all the data from the official government archive[4]. Table 1 summarizes the statistics of the two corpora, tokenization was done at the word-level.

We split the documents into individual sentences, and created batches that were then distributed to the annotators each week in groups of three. We continued this process until the annotation period was finished, concluding with 60,000 annotated sentences. The PolitiCAUSE corpus is the compilation of the annotated sentences that met data quality standards.

4.2. The Annotation Process

We decided to exclude crowd sourced platform annotators for this task after preliminary tests showed low quality annotations. This decision is in line with the findings of previous studies (Tan et al., 2023; Dunietz et al., 2017). We hired 12 political science graduate students from an international university to annotate the data, and were compensated in accordance to the country's research pay scale. Annotator demographics range from different regions of the world, including underrepresented regions, as well as gender. The annotation process lasted for five months and used Spacy's Prodigy (Montani and Honnibal) platform, which was deployed on an AWS instance. The

[4]https://www.gov.uk/search/news-and-communications

annotators accessed the annotation task from their own personal computers, and annotated freely after training was concluded. Figure 1 shows an annotation example from our codebook. We trained the annotators in three iterations, each followed by a feedback session. Annotators were also encouraged to communicate issues that may arise on our working channel. We communicated that they were working with political discourse that may have false representations of the world, or unethical political stances, and that in no way represented any of the researchers' or the annotators' political views.

Figure 1: Prodigy annotation example with a toy sentence.

4.3. Statistics

PolitiCAUSE is composed of 17,780 unique sentences that were annotated by at least two human coders, producing a total of 55,754 annotated samples. We pruned the PolitiCAUSE dataset to ensure data quality: removing all data points that had fewer than 2 annotations per sentence, excluding sentences that did not have a clear majority agreement for their label, and excluding sentences that had a mean confidence score lower than 3. Section 9.1 includes examples of annotated sentences.

In total 12,710 (71%)sentences were assigned a "not causal" label, and 5,070 (29%) a "causal" label, making our dataset imbalanced, which is expected on this task. Most sentence were annotated at least three times, and we used majority voting to assign labels. We assigned each sentence either a positive class "causal" (1) or a negative class "not causal" (0). When a clear majority could not be reached, we discarded the sentences (values between 0.4 and 0.6 in the ratio of causal to non causal assigned label).

We calculated the mean self-reported confidence score for each sentence by averaging the confidence scores each coder reports per annotation. The confidence score ranges from 0 to 5. We removed sentences with a mean confidence score lower than 3 to ensure high quality data. The mean value for our coders is of 4.49 after removing the values less than 3. The self-reported confidence score is higher for non causal sentences (4.63) than for causal sentences (4.13)[5]. This difference suggests that causal sentences are more complex and nuanced than non-causal sentences, hence harder to annotate.

	Total	Not Causal	Causal
train	12446	8897	3549
val	2667	1906	761
test	2667	1907	760

Table 2: Dataset splits used in the benchmark. They follow a 70%-15%-15% split for training, validation and inference.

4.4. Benchmark

To evaluate the usefulness and difficulty of PolitiCAUSE, we also present a benchmarking study using transformer-based models. We compare the performance of three classification models, using various evaluation metrics. To this end, we divided our dataset into three subsets: training (70%), validation (15%) and testing (15%). We ensured that the splits have a similar distribution of the label class as the original dataset. Figure 2 describes the statistics.

We used three popular BERT-based models:

- The Bi-directional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) model, which has an "encoder-only" transformer architecture, with multiple self-attention heads, trained via Masked Language Modeling and Next Sentence Prediction tasks.
- The Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019) model, which has the same architecture as BERT, but eliminates the next sentence prediction task during pre-training, and utilizes dynamic masking technique.
- DistilBERT (Sanh et al., 2019) that was created by applying knowledge distillation to the BERT model, eliminating token-type embeddings and the pooler from the architecture, and reducing in half the number of layers.

We followed the standard procedure for NLP classification tasks, and used Hugging face's base tokenizer and configuration for each model[6][7][8]. Using the train and validation splits of our data, we trained for 10 epochs and saved the best epoch as the fine-tuned model for inference, see Section 9.3 for further information on parameter settings. We used 1 NVIDIA A100 40GB HBM2 GPU for the experiment, which took less than 2 total GPU hours to complete. To analyze results we included the standard metrics plus the

[5] t-value = 68.58, p-value = 0.00

[6] https://github.com/huggingface/transformers/tree/v4.32.1/src/transformers/models/bert

[7] https://github.com/huggingface/transformers/blob/v4.32.1/src/transformers/models/roberta

[8] https://github.com/huggingface/transformers/blob/v4.32.1/src/transformers/models/distilbert

	BERT	RoBERTa	DistilBERT	UniCausal
Acc	0.832	0.836	0.832	0.715
Prec	0.671	0.686	0.696	0.500
Recall	0.805	0.783	0.730	0.612
MCC	0.617	0.617	0.594	0.550
F1	0.732	0.731	0.712	0.349

Table 3: Results fine-tuned models

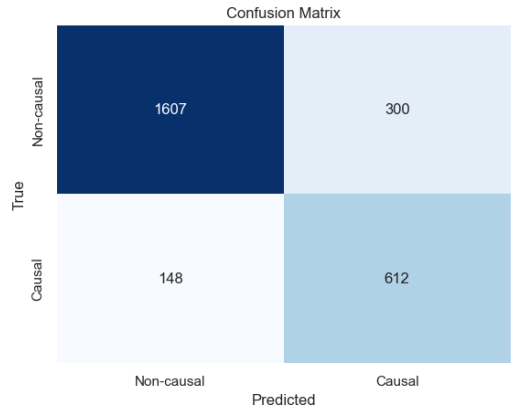

Figure 2: A confusion matrix for inference with fine-tuned BERT; For all the model evaluation results, see Section 9.4.

Matthews Correlation Coefficient, which is more suitable for imbalanced classes (Boughorbel et al., 2017). With the test split, we then used the fine-tuned models for inference. To to analyze our results against a specialized model, we included UniCausal in our inference. UniCausal (Tan et al., 2022) is a BERT-based model finetuned using the UniCausal dataset.

Table 3 shows the results of all four models. Regarding the fine-tuned models the results are promising, showing a high level of accuracy and precision. Furthermore, a MCC score around 0.60 shows that the classifiers are able to distinguish between the two classes moderately well, even if the dataset is imbalanced. Furthermore, the results of the UniCausal model show us that political text is distinct from other genres, and that a specialized corpus leads to higher results. The best overall model is the BERT-based model, with an F1 score of 0.73, only marginally higher than the RoBERTa model. Even if we see a high accuracy across the models (0.83), we can see that precision is 0.10 points lower across the board.

4.5. Error Analysis

We undertook an error analysis to gain deeper insights of the results. First, we identified sentences that were consistently misclassified across all three fine-tuned models (excluding UniCausal), and created an error subset. This process yielded 248 sentences, with 157 (63%) causal and 91 (37%) non-causal instances. Directly, this indicates an overestimation of causal sentences, a significant finding given the imbalanced nature of the dataset.

	Full corpus		Error subset	
Label	0 (TN)	1 (TP)	0 (FN)	1 (FP)
Mean conf	4.63	4.13	4.20	4.30
Maj label	0.79	0.10	0.71	0.26
Mean len	26.13	31.73	30.10	31.35
Causal conn	0.37	0.50	0.44	0.48

Table 4: Error analysis statistics

Next, we compared corpus statistics with the error subsets'. The mean confidence score of the full corpus is 4.49 (not causal = 4.63, causal = 4.13); for the error subset, the mean confidence is 4.26 (false negative = 4.20, false positive = 4.30). There is a lower mean confidence score in the error subset compared to the full corpus; the false positive value is higher than the false negative value, which is unexpected, given that their based on their true labels we would expect the opposite.

We continued the analysis with the majority label ratio. The corpus had an average ratio of 0.10 for non-causal samples and of 0.79 for causal samples [9]. The error subset statistics are of 0.71 for false negatives, and 0.26 for the false positives. This difference indicate that the mislabeled sentences had greater annotator disagreement.

We also evaluated sentence length, given that multiple events in a single sentence or richer descriptions increase annotation difficulty. The difference between the mean sentence length between non causal (26.13) and causal sentences (31.73) in the corpus was statistically significant[10]. In the error analysis subset (false negatives - 30.1; false positives - 31.35) the difference was not statistically significant[11].

Finally, we examined content. Even though we do not expect causal connectors in non-causal sentences, occasionally, these expressions are used to indicate relationships other than causality; Moreover, causal connectors can be present as part of an incomplete structure (see Example 5 in Section 9.1), leading to sentences being labeled as 0. Using causal expressions identified in the literature (Mirza et al., 2016), we compared their presence in the data. In the full corpus, causal connectors were only present in 37% of the non causal data, while 50% of the causal instances included such a connector. However, in the error subset, 44% of false negatives sentences and 48% of false positives included a causal signal.

All together, this error analysis indicates that the mislabeled sentences do not follow the patterns expected from the corpus: they were harder to annotate in average and generated more disagreement across annotators. Potentially, this could be due to sentence

[9] Sentences with a value between between 0 - 0.4, were label as 0; between 0.6 - 1, as 1; Values closer to 0 or 1 mean higher agreement between annotators.

[10] t=21.40, p-value=0.0

[11] t=-0.66, p-value=0.51

complexity or causal ambiguity.

5. Conclusion

In this paper, we presented PolitiCAUSE, a new corpus of political sentences taken from the UNGDC and UKPress corpora and annotated for causality. We developed an annotation scheme, which we rolled out to 12 participants who annotated the corpus with two types of information: (1) whether a sentence contains a causal relation or not, and (2) the spans of text that correspond to the cause and effect components of the causal relation. We developed an annotation scheme that underwent several iterations and revisions to ensure its quality and reliability. We also provided detailed statistics and analysis of the corpus. We finished by included an error analysis section. The main limitations of the dataset include the lack of annotation across sentences for speeches and the absence of implicit causality. These will be addressed in future work.

Furthermore, we conducted a benchmarking study using three Bert-based classifiers on PolitiCAUSE. We compared the performance of all BERT-based models, using various evaluation metrics. Models achieved a moderate performance on the dataset, with a MCC score between 0.59 - 0.62, showing that there is room for improvement and perhaps additional revision. Additionally, we compared the results to a specialized model, and consider political text to be sufficiently distinct from other genres to warrant its own annotation. In future work, we are looking to include LLMs to compare human and machine annotation of causality, and introduce implicit causality and inter-sentential relationships.

We hope that PolitiCAUSE will encourage further research on causality in texts, especially in the domain of political debate. We believe that understanding causality in texts can help to enhance various natural language processing applications, such as explanation generation, summarization, question answering, argumentation mining, and sentiment analysis. We also believe that understanding causality in political texts can help address complex problems in political communication analysis, as well as in health and climate change policy, where we need to identify the factors that contribute to or prevent diseases and environmental changes, and evaluate the effects of different policy interventions. We welcome feedback and suggestions from the research community on how to improve and extend the corpus. We also encourage researchers to use PolitiCAUSE for their own experiments and projects on causality in texts.

6. Ethics Statement

Like any training set, our data has inherent biases. Causal language used can reflect underlying mental models that may be racist, sexist, xenophobic, or derogatory towards specific groups. Additionally our dataset includes a variety of political ideologies and national perspectives, but dominant viewpoints are more likely to take precedence over alternative perspectives. It's crucial to differentiate between identifying these expressions, and endorsing them. The capability to automatically identify the causal connections established in political communication allows for more efficient detection of false information, hate speech, and harmful content, which is our objective.

7. Acknowledgements

The authors thank the DFG (EXC number 2055 – Project number 390715649, SCRIPTS) for providing funding for the annotation efforts. This project has also received funding from the European Union's Horizon Europe research and innovation program under Grant Agreement No 101057131, Climate Action To Advance HeaLthY Societies in Europe (CATALYSE).

8. Bibliographical References

Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE*, 12(6):e0177678.

Tommaso Caselli and Piek Vossen. 2017. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [cs].

Brett Drury, Hugo Gonçalo Oliveira, and Alneu de Andrade Lopes. 2022. A survey of the extraction and applications of causal relations. *Natural Langauge Engineering*.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a New Dataset for Exploring Explainable Causal Reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE Corpus 2.0: Annotating Causality and Overlapping Relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2022. Scaling up Discourse Quality Annotation for Political Science. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. SemEval-2007 Task 04: Classification of Semantic Relations between Nominals. In *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 13–18.

Jinghang Gu, Longhua Qian, and Guodong Zhou. 2016. Chemical-induced disease relation extraction with various linguistic features. *Database*, page 42.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Christopher Hidey and Kathy McKeown. 2016. Identifying Causal Relations Using Parallel Wikipedia Articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany. Association for Computational Linguistics.

Slava Jankin, Alexander Baturo, and Niheer Dasandi. 2023. United Nations General Debate Corpus.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues de Mazancourt, and Mahmoud El-Haj. 2021. The financial document causality detection shared task (FinCausal 2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 58–60, Lancaster, United Kingdom. Association for Computational Linguistics.

Paramita Mirza, Sara Tonelli Fondazione, and Bruno Kessler. 2016. CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 64–75, Osaka, Japan.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Ines Montani and Matthew Honnibal. Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models.

Ad Neeleman and Hans Van de Koot. 2012. The Linguistic Expression of Causation. In *The Theta System: Argument Structure at the Interface*. Oxford University Press, Oxford.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, and Aravind Joshi. 2007. The Penn Discourse Treebank 2.0 Annotation Manual. Publication Title: IRCS Technical Reports Series.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn discourse treebank version 3.0.

Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. 2018. Feasible Annotation Scheme for Capturing Policy Argument Reasoning using Argument Templates. In *Proceedings of the 5th Workshop on Argument Mining*, pages 79–89, Brussels, Belgium. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Steven Sloman. 2005. *Causal Models: How People Think about the World and Its Alternatives*. Oxford University Press.

Torgrim Solstad and Oliver Bott. 2017. 619Causality and Causal Reasoning in Natural Language. In *The Oxford Handbook of Causal Reasoning*. Oxford University Press.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022. The Causal News Corpus: Annotating Causal Relations in Event Sentences from News. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Fiona Anting Tan, Xinyu Zuo, and See-Kiong Ng. 2023. UniCausal: Unified Benchmark and Repository for Causal Text Mining. ArXiv:2208.09163 [cs].

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2014. Tempeval-3: Evaluating events, time expressions, and temporal relations.

Konstantin Vössing. 2023. Argument-stretching: (slightly) invalid political arguments and their effects on public opinion. *European Political Science Review*, pages 1–21. Publisher: Cambridge University Press.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A Survey on Extraction of Causal Relations from Natural Language Text. *Knowledge and Information Systems*.

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting Causal Language Use in Science Findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4664–4674, Hong Kong, China.

9. Appendix

9.1. Annotated Examples

In this section we provide 6 examples with descriptions. We included 3 causal and 3 non-causal sentences, sampled from the UNGD Corpus and the UK-Press corpora. First, we introduce the sentence. Below each sentence, is the label and a description.

1. **Sentence:** 1,275 Azerbaijanis were taken hostage, out of which 150 are still missing.

   Label: Not Causal.

   Description: There are no events causally linked in the sentence.

2. **Sentence:** International solidarity also has the power to prevent a climate disaster.

   Label: Causal

   Description: Event 1 “International Solidarity”, in green, is tagged as the cause of Event 2 “climate disaster” which is tagged as the effect, in yellow.

3. **Sentence:** 2 million doses from the Government of the United States of America — our main trading partner.

   Label: Not Causal

   Description: There are no events causally linked in the sentence.

4. **Sentence:** A divergence of views on the nuclear deal with Iran generated the current tensions.

   Label: Causal

   Description: Event 1 “A divergence of views on the nuclear deal”, in green, is tagged as the cause of Event 2 “the current tensions”, which is tagged as the effect, in yellow.

5. **Sentence:** To combat terrorism and transborder crime, as well as enhancing cybersecurity.

   Label: Not Causal

   Description: This is an incomplete causal statement, with a missing “cause” event.

6. **Sentence:** COVID-19 has triggered the most severe recession in almost a century.

   Label: Causal

   Description: Event 1 “COVID-19”, in green, is tagged as the cause of Event 2 “most sever recession”, which is tagged as the effect, in yellow.

9.2. Examples of three annotations of a single sentence

In this section we provide 2 examples of span annotations. First we introduce the sentence, then three annotations and a short description.

1. **Sentence:** COVID-19 has triggered the most severe recession in almost a century.

   A_1: COVID-19 has triggered the most severe recession in almost a century.

   A_2: COVID-19 has triggered the most severe recession in almost a century .

   A_3: COVID-19 has triggered the most severe recession in almost a century.

   Description: There is full agreement on Event 1 "COVID-19" as the "cause". However, there is disagreement on how much should the "effect" span include.

We added the "subject" tag in the annotation guidelines to help annotators identify "cause" and "effect" events with higher precision (see Section 3.2). The next example is a sentence with three distinct annotations, that contain a "subject" tag:

1. **Sentence:** France's position enabled the deal to be completed.

   A_1: France's position enabled the deal to be completed .

   A_2: France's position enabled the deal to be completed .

   A_3: France's position enabled the deal to be completed .

   Description: There is full agreement over the "effect" tag, and almost full agreement over the "cause" tag; However, adding "France" as an subject tag, increases precision of the span's length and improves agreement across coders.

9.3. Hyperparameter Specifications for the benchmark models

We used the PyTorch-Transformers library, a comprehensive toolkit with implementations, pre-trained model weights, usage scripts, and conversion utilities. It's standardized features ensured consistency and reproducibility throughout the benchmark. All fine-tuning models used during the benchmark in our study adhered to the same parameters, upholding uniformity and ensuring equitable evaluation criteria.

	All models
max length	512
learning rate	2e5
train batch size	16
val batch size	16
num train epochs	10
weight decay	0.01
eval strategy	epoch
max length	512

Table 5: Parameter setting. Saving strategy was based on best epoch.

9.4. Confusion matrices used during evaluation.

Inference results of the RoBERTa and DistilBERT fine-tuned models on the PolitiCAUSE data. We can observe similar distribution of the classes, also closely matching the BERT model results, in Fig. 2.

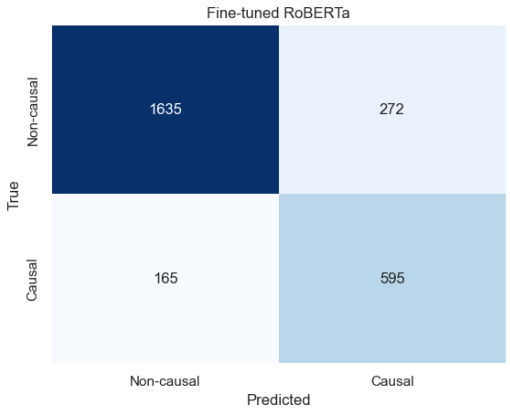

Figure 3: Results RoBERTa

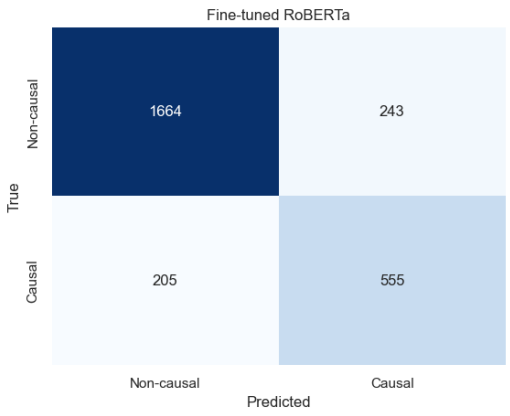

Figure 4: Results DistilBERT.