UniCausal: Unified Benchmark and Repository for Causal Text Mining

DAWAK 2023, 28 – 30 August 2023, Penang, Malaysia

Fiona Anting Tan, Xinyu Zuo and See-Kiong Ng

https://github.com/tanfiona/UniCausal

https://link.springer.com/chapter/10.1007/978-3-031-39831-5_23

Extracting information about causal relationships from text have important applications in NLP.

- Causal text mining relates to the extraction of causal information from text. Given an input sequence, we are interested to know if and where causal information occurs.
- Extracted causal information is useful for various downstream NLP challenges like: summarization, prediction, natural language understanding, etc.

Constructing and interpreting causal knowledge graphs from news. (Tan et al., AAAI-SS 2023)

Panasonic

NUS National University of Singapore

The Tata Avinya concept car is unveiled during a global launch

Fifteen countries have agreed to work together toward 100% zero-emission new truck and bus sales by 2040. Participants in this coordinated global effort agree that zero-emission trucks and buses are essential to reducing transport emissions, mitigating climate change, improving air quality, reducing the use of fossil fuels and energy costs. The Global Agreement on Zero-Emission Trucks and Buses builds on previous action and ambition from key government ...

News Articles

Extraction

Cause

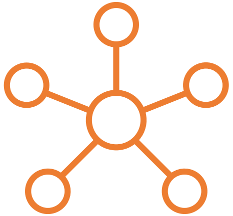
the chip crisis and the surge in Covid cases in some parts of the world

Effect

Global car manufacturers like Toyota Motor Corp and Ford Motor Corp have decided to slash production sharply

Causal Relations

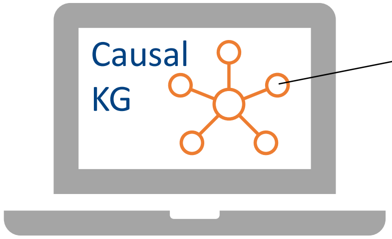
Clustering & Representation



Causal Knowledge Graph

Fiona Anting Tan, Debdeep Paul, Sahim Yamaura, Miura Koji, and See-Kiong Ng. 2023. Constructing and Interpreting Causal Knowledge Graphs from News. In *Proceedings of the AAAI Summer Symposium on AI for FinTech, 2023.*

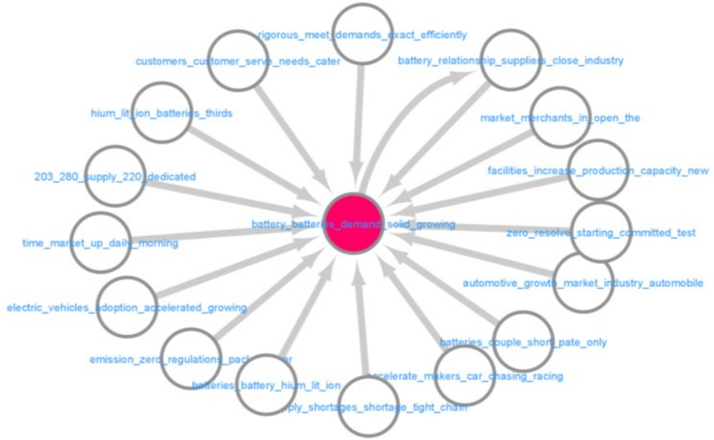
Parts Purchaser for EV Batteries



1. Purchaser can search for relevant nodes to understand about the factors that affect battery demand from KG

2. Purchaser can set these *Cause* topics as new alerts

Search Node:
battery_batteries_demand_solid_growing



Identified Causes:

- *fuel_environmental_concerns_increasing_costs:* The increasing high fuel costs and environmental concerns
- *battery_relationship_suppliers_close_industry:* together with the TESLA project … promoting as part of the European battery projects
- *electric_vehicles_adoption_accelerated_growing:* the electric vehicle revolution

…

Fiona Anting Tan, Debdeep Paul, Sahim Yamaura, Miura Koji, and See-Kiong Ng. 2023. Constructing and Interpreting Causal Knowledge Graphs from News. In *Proceedings of the AAAI Summer Symposium on AI for FinTech, 2023.*

What does it mean to design a model that successfully extracts causal relations from text?

- Type of annotations
- Domain covered
- Exclusion rules

…

Corpus	*Causal* Example
AltLex	<ARG0>In the Philippines , Washi</ARG0> caused <ARG1>at least 1,268 deaths .</ARG1>
BECAUSE	<ARG0>Having only a Republican measure</ARG0> makes <ARG1>the task harder</ARG1>.
CTB	Iraq said it <ARG1>invaded</ARG1> Kuwait because of <ARG0>disputes</ARG0> over oil and money.
ESL	Ten <ARG1>dead</ARG1> in southern Iran <ARG0>quake</ARG0>.
PDTB	<ARG1>And the firms are stretching their nets far and wide</ARG1> <ARG0>to do it</ARG0>.
SemEval	The front <ARG0>wheels</ARG0> are making a <ARG1>grinding noise</ARG1> .

Table 1. Example data from the six causal text mining corpora.

Corpus	Source	Inter-sent	Linguistic	Arguments
AltLex (Hidey and McKeown, 2016)	News		AltLex	Words before/after signal
BECAUSE 2.0 (Dunietz et al., 2017b)	News, Congress Hearings		Explicit	Phrases
CausalTimeBank (CTB) (Mirza et al., 2014)	News	✓	All	Event head word(s)
EventStoryLine V1.0 (ESL) (Caselli and Vossen, 2017)	News	✓	All	Event head word(s)
Penn Discourse Treebank V3.0 (PDTB) (Webber et al., 2019)	News	✓	All	Clauses
SemEval 2010 Task 8 (SemEval) (Hendrickx et al., 2010)	Web		All	Noun phrases

Table 2. Properties of six causal text mining corpora.

UniCausal focuses on three tasks in causal text mining.

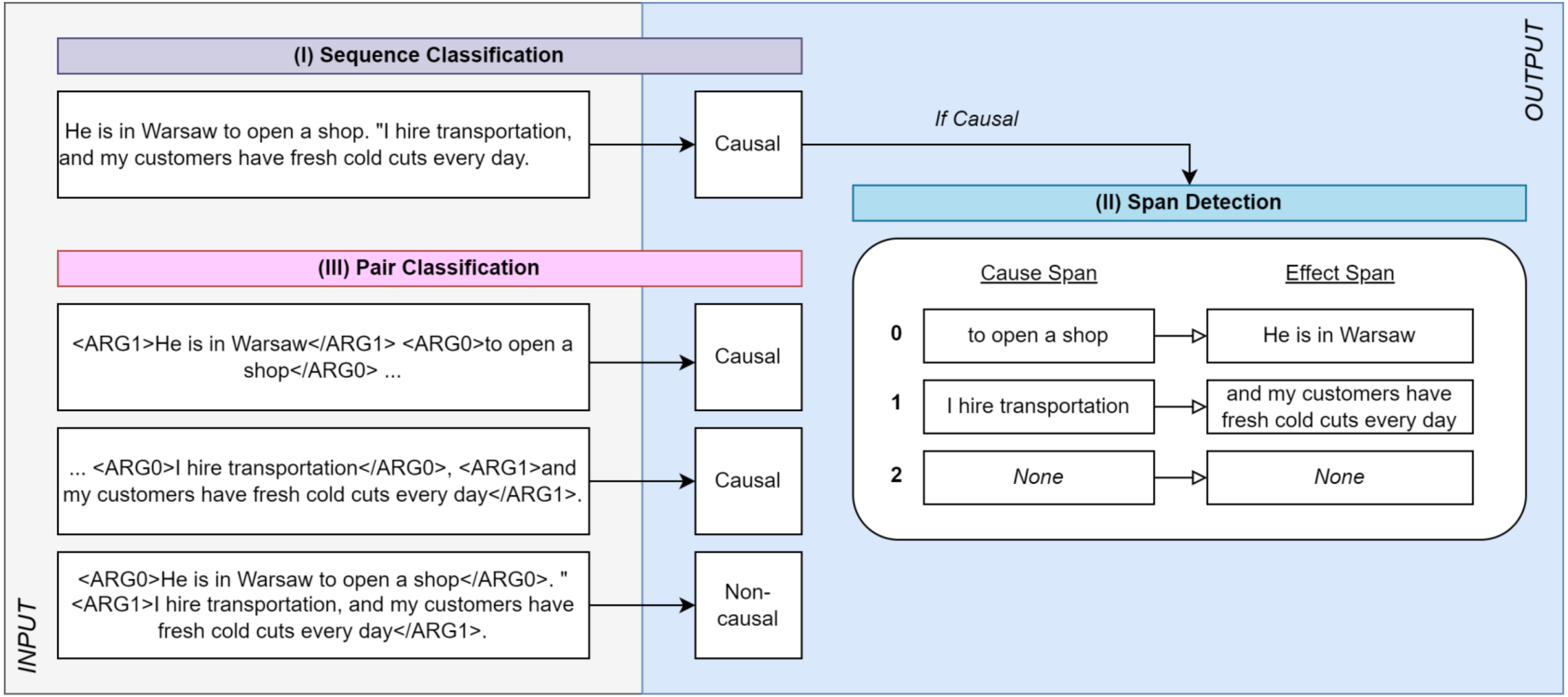


Figure 1. A two-sentence example that contains causal relations.

We formatted six datasets into a consistent format for causal text mining.

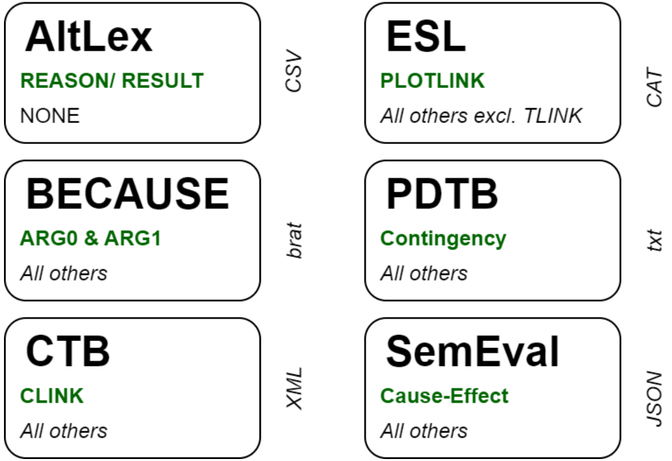


Figure 2. Annotation convention equivalent to *Causal* (in green) and data format (right of box) of the six causal text mining corpora

Input

(I) Seq

(II) Span

corpus	because
doc_id	Article247_327.ann
sent_id	3
eg_id	1
index	because_Article247_
text	They will then score one point for every subsequent issue or broadcast or Internet posting after the first offense is noted by Chatterbox if they continue not to report said inconvenient fact--and an additional two points on days when the news organization runs a follow-up without making note of said inconvenient fact.
seq_label	1
num_sents	1
causal_text_w_pairs	['<ARG1>They will then score one point for every subsequent issue or broadcast or Internet posting after the first offense is noted by Chatterbox</ARG1> if <ARG0>they continue not to report said inconvenient fact</ARG0>--and an additional two points on days when the news organization runs a follow-up without making note of said inconvenient fact.', '<ARG1>They will then score one point</ARG1> for <ARG0>every subsequent issue or broadcast or Internet posting after the first offense is noted by Chatterbox</ARG0> if they continue not to report said inconvenient fact--and an additional two points on days when the news organization runs a follow-up without making note of said inconvenient fact.']
num_rs	2

Table 3. Example data in UniCausal.

We formatted six datasets into a consistent format for causal text mining.

corpus	because
doc_id	Article247_327.ann
sent_id	3
eg_id	1
index	because_Article247_327.ann_3_1
text	They will then score one point for every subsequent issue or broadcast or Internet posting after the first offense is noted by Chatterbox if they continue not to report said inconvenient fact--and an additional two points on days when the news organization runs a follow-up without making note of said inconvenient fact.
text_w_pairs	<ARG1>They will then score one point for every subsequent issue or broadcast or Internet posting after the first offense is noted by Chatterbox</ARG1> if <ARG0>they continue not to report said inconvenient fact</ARG0>--and an additional two points on days when the news organization runs a follow-up without making note of said inconvenient fact.
seq_label	1
pair_label	1
context	
num_sents	1

corpus	because
doc_id	Article247_327.ann
sent_id	3
eg_id	2
index	because_Article247_327.ann_3_2
text	They will then score one point for every subsequent issue or broadcast or Internet posting after the first offense is noted by Chatterbox if they continue not to report said inconvenient fact--and an additional two points on days when the news organization runs a follow-up without making note of said inconvenient fact.
text_w_pairs	<ARG1>They will then score one point</ARG1> <ARG0>every subsequent issue or broadcast or Internet posting after the first offense is noted by Chatterbox</ARG0> if they continue not to report said inconvenient fact--and an additional two points on days when the news organization runs a follow-up without making note of said inconvenient fact.
seq_label	1
pair_label	1
context	
num_sents	1

Input

(III) Pair

Table 4. Example ungrouped version of UniCausal, data is equivalent to Table 3.

We formatted six datasets into a consistent format for causal text mining.

- Train-test splits: Followed original corpora recommendations

- Limitations: We restricted our repository to examples that are <=3 sentences-long and contain <=3 causal relations.

Corpus	Split	(I) Seq *Non-causal*	(I) Seq *Causal*	(II) Span *Causal*	(III) Pair *Non-causal*	(III) Pair *Causal*
AltLex	Train	277	300	300	296	315
	Test	286	115	115	289	127
BEC-AUSE	Train	183	716	716	266	902
	Test	10	41	41	14	46
CTB	Train	1,651	234	-	3,047	270
	Test	274	42	-	444	48
ESL	Train	957	1,043	-	-	-
	Test	119	113	-	-	-
PDTB	Train	24,901	8,917	8,917	32,587	9,809
	Test	5,796	2,055	2,055	7,694	2,294
Sem-Eval	Train	6,976	999	-	6,997	1,003
	Test	2,387	328	-	2,389	328
Total		43,817	14,903	12,144	54,023	15,142

Table 5. UniCausal's data sizes split by corpus source and task.

We wrote a custom `load_cre_dataset` function so that users can work with the data directly by calling the data name.

```
In [2]:
from _datasets.unifiedcre import load_cre_dataset, available_datasets
print('List of available datasets:', available_datasets)

"""
 Example case of loading AltLex and BECAUSE dataset,
 without adding span texts to seq texts, span augmentation or user-provided datasets,
 and load both training and validation datasets.
"""
load_cre_dataset(dataset_name=['altlex','because'], do_train_val=True, data_dir='../data')
```

```
List of available datasets: ['altlex', 'because', 'ctb', 'esl', 'esl2', 'pdtb', 'semeval2010t8', 'cnc', 'causenet', 'c
ausenetm']
```

```
Out[2]: (DatasetDict({
            span_validation: Dataset({
                features: ['corpus', 'index', 'text', 'label', 'ce_tags', 'ce_tags1', 'ce_tags2'],
                num_rows: 156
            })
            span_train: Dataset({
                features: ['corpus', 'index', 'text', 'label', 'ce_tags', 'ce_tags1', 'ce_tags2'],
                num_rows: 1016
            })
        }),
        DatasetDict({
            seq_validation: Dataset({
                features: ['corpus', 'index', 'text', 'label'],
                num_rows: 296
            })
            pair_validation: Dataset({
                features: ['corpus', 'index', 'text', 'label'],
                num_rows: 476
            })
            seq_train: Dataset({
                features: ['corpus', 'index', 'text', 'label'],
                num_rows: 460
```

Figure 3. Screenshot of tutorial on Github

`load_cre_dataset` formats all the examples into inputs that can be fed directly into Huggingface pipelines.

(I) Seq


```
{'corpus': 'altlex', 'index': 'altlex_altlex_dev.tsv_0_0', 'text': "The Bhopal disaster , also referred to as the Bhop
al gas tragedy , was a gas leak incident in India , considered the world 's worst industrial disaster .", 'label': 0}
```


(II) Span


```
{'corpus': 'altlex', 'index': 'altlex_altlex_dev.tsv_4_0', 'text': ['The', 'U.S.', 'Supreme', 'Court', 'refused', 't
o', 'hear', 'an', 'appeal', 'of', 'the', 'decision', 'of', 'the', 'lower', 'federal', 'courts', 'in', 'October', '199
3', ',', 'meaning', 'that', 'victims', 'of', 'the', 'Bhopal', 'disaster', 'could', 'not', 'seek', 'damages', 'in',
'a', 'U.S.', 'court', '.'], 'label': 1, 'ce_tags': ['B-C', 'I-C', 'I-C', 'I-C', 'I-C', 'I-C', 'I-C', 'I-C', 'I-C', 'I-
C', 'I-C', 'I-C', 'I-C', 'I-C', 'I-C', 'I-C', 'I-C', 'I-C', 'I-C', 'I-C', 'I-C', 'O', 'B-E', 'I-E', 'I-E', 'I-E', 'I-
E', 'I-E', 'I-E', 'I-E', 'I-E', 'I-E', 'I-E', 'I-E', 'I-E', 'I-E', 'I-E'], 'ce_tags1': ['O', 'O', 'O', 'O', 'O', 'O',
'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O',
'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], 'ce_tags2': ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O',
'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O',
'O']}
```


(III) Pair


```
{'corpus': 'altlex', 'index': 'altlex_altlex_dev.tsv_0_0', 'text': "<ARG1>The Bhopal disaster , also referred to</ARG1
> as <ARG0>the Bhopal gas tragedy , was a gas leak incident in India , considered the world 's worst industrial disast
er .</ARG0>", 'label': 0}
```

For our baseline models, we fine-tuned pretrained BERT.

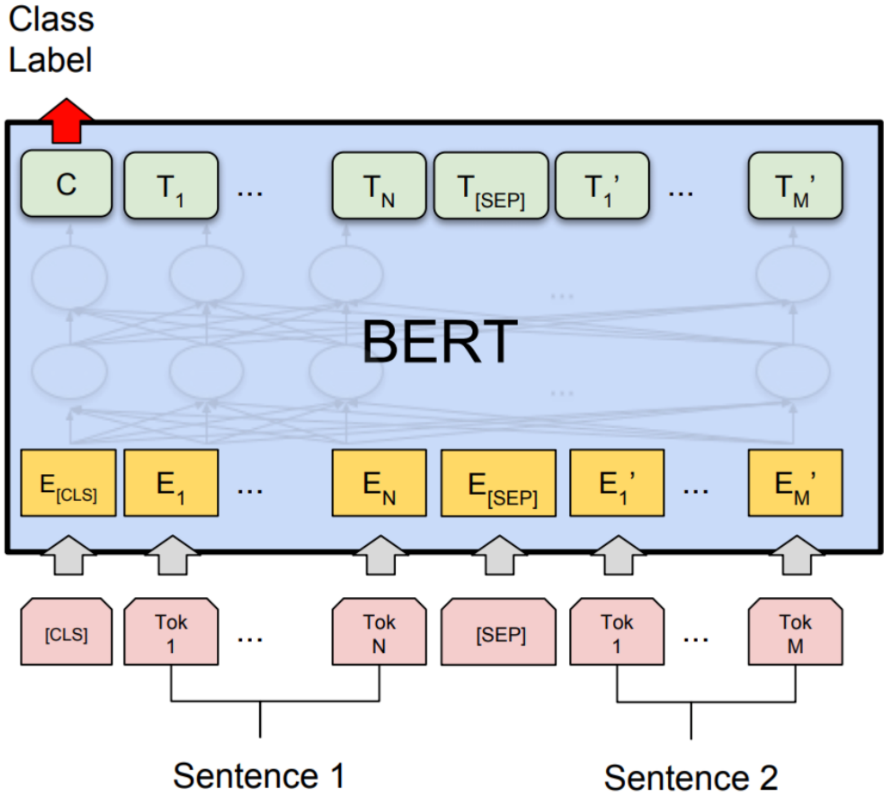


Figure 4: BERT for Sequence Classification (Devlin et al., 2018)

For our baseline models, we fine-tuned pretrained BERT.

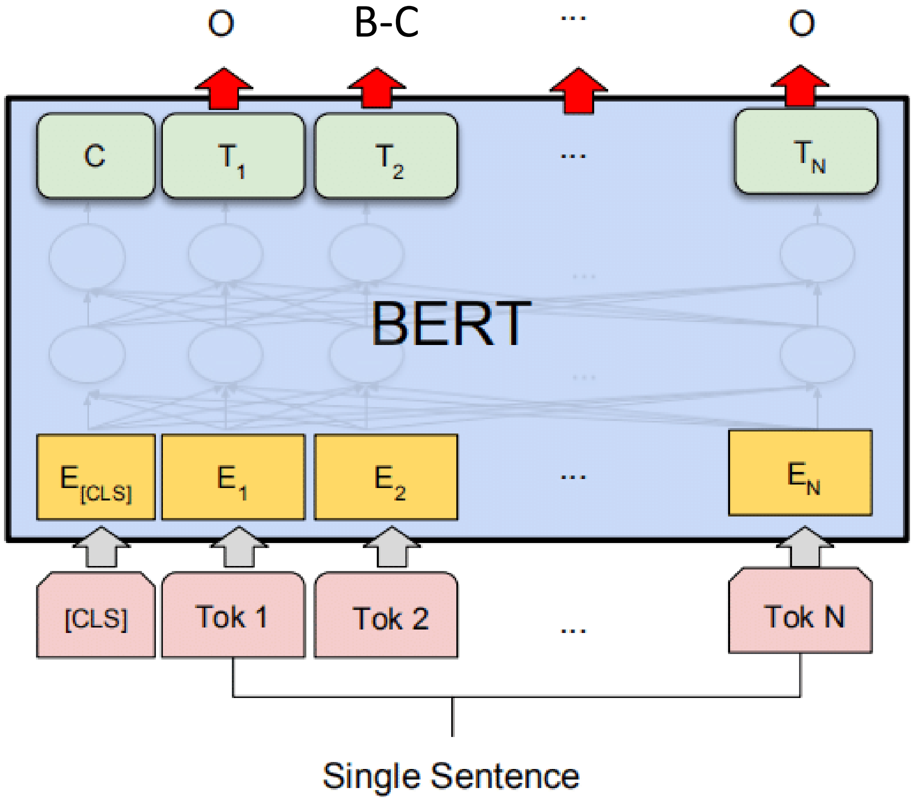
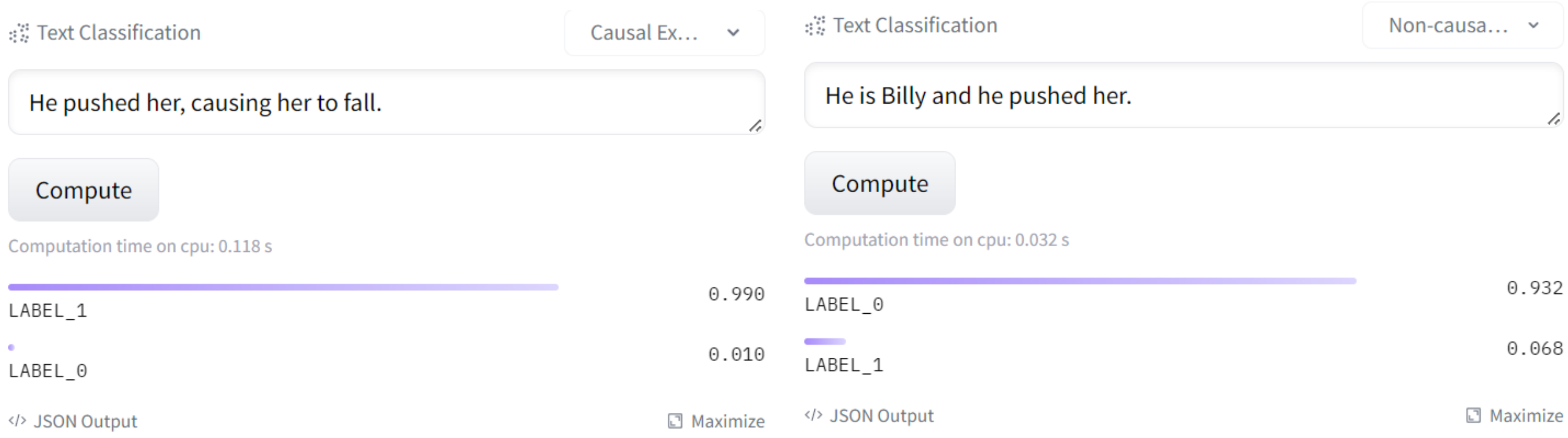
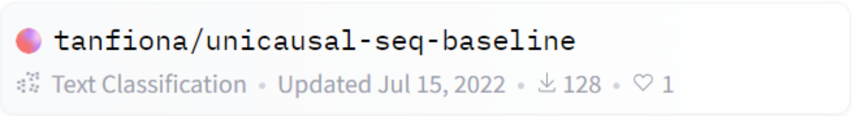


Figure 5: BERT for Token Classification (Devlin et al., 2018)

Our baseline models are available on Huggingface Hub.





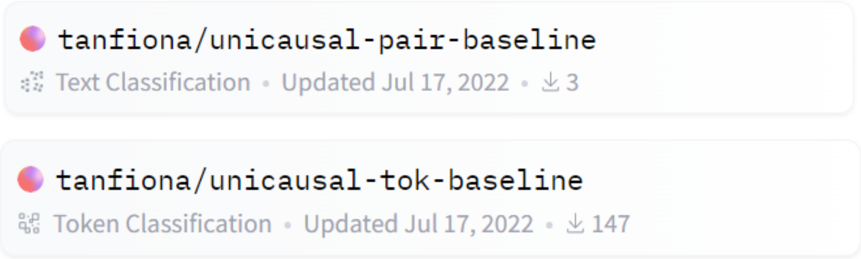


Figure 6. Pretrained model checkpoints on Huggingface Hub

We provide initial baseline scores across datasets and tasks for the causal text mining community to beat.

Test Set	(I) Sequence Classification				(II) Span Detection			(III) Pair Classification			
	P	R	F1	Acc	P	R	F1	P	R	F1	Acc
All	71.13 ±0.80	69.14 ±1.60	70.10 ±0.58	86.27 ±0.15	46.33 ±1.22	60.35 ±0.30	52.42 ±0.90	85.44 ±0.96	83.93 ±0.44	84.68 ±0.27	93.68 ±0.16
AltLex	50.76 ±1.61	63.48 ±4.60	56.37 ±2.49	71.87 ±1.19	27.74 ±1.20	42.99 ±0.85	33.72 ±1.12	82.60 ±1.99	87.09 ±1.53	84.76 ±0.66	90.43 ±0.55
BECAUSE	92.32 ±1.69	70.24 ±2.04	79.77 ±1.68	71.37 ±2.24	32.51 ±2.82	44.30 ±2.33	37.47 ±2.57	87.93 ±1.73	94.78 ±1.94	91.21 ±1.18	86.00 ±1.90
CTB	42.37 ±2.11	66.19 ±4.26	51.58 ±1.82	83.48 ±1.21	-	-	-	75.66 ±3.61	72.50 ±6.81	73.94 ±4.68	95.04 ±0.78
ESL	76.11 ±2.04	67.43 ±3.45	71.45 ±1.89	73.79 ±1.34	-	-	-	-	-	-	-
PDTB	72.59 ±0.61	66.34 ±1.63	69.31 ±0.70	84.63 ±0.17	47.77 ±1.22	61.54 ±0.29	53.78 ±0.88	84.56 ±1.17	82.04 ±0.46	83.28 ±0.36	92.43 ±0.23
SemEval	73.39 ±1.18	89.51 ±1.59	80.64 ±0.46	94.81 ±0.16	-	-	-	93.38 ±0.88	96.10 ±0.59	94.71 ±0.23	98.70 ±0.07

Table 6. Mean and standard deviation of performance metrics for different test sets across the three tasks. All models were trained on all six datasets, where applicable.

We compare the F1 scores when training and testing on different corpus to review cross-corpora compatibility.

(I) Sequence Classification

	Test Set						
Training Set	All	AltLex	BECAUSE	CTB	ESL	PDTB	SemEval
All	**70.10 ±0.58**	**56.37 ±2.49**	79.77 ±1.68	51.58 ±1.82	71.45 ±1.89	69.31 ±0.70	80.64 ±0.46
AltLex	32.93 ±3.57***	51.85 ±2.53	36.47 ±11.18***	38.21 ±6.20*	53.30 ±8.37**	22.91 ±5.79***	55.83 ±6.68***
BECAUSE	39.15 ±0.99***	47.02 ±1.52**	**90.77 ±2.22*** **	25.17 ±1.34***	63.49 ±1.94**	42.49 ±0.68***	23.71 ±1.93***
CTB	33.49 ±5.48***	55.91 ±7.63	54.73 ±9.40**	**63.65 ±5.55** **	33.26 ±15.44**	25.97 ±3.73***	51.76 ±13.85**
ESL	39.62 ±0.89***	46.29 ±1.15**	90.12 ±1.05***	30.84 ±1.35***	**81.21 ±2.35*** **	42.55 ±1.25***	26.15 ±2.62***
PDTB	60.99 ±0.76***	48.94 ±1.88**	69.61 ±2.16**	39.54 ±1.88***	38.71 ±3.15***	**70.31 ±0.56* **	19.75 ±3.35***
SemEval	28.25 ±0.86***	28.95 ±1.74***	16.91 ±3.40***	38.51 ±3.44**	45.95 ±3.50***	10.11 ±1.61***	**89.58 ±0.71*** **

(II) Span Detection

	Test Set			
Training Set	All	AltLex	BECAUSE	PDTB
All	**52.42 ±0.90**	**33.72 ±1.12**	37.47 ±2.57	53.78 ±0.88
AltLex	6.20 ±0.74***	21.45 ±1.87***	11.51 ±1.63***	5.47 ±0.76***
BECAUSE	12.74 ±0.35***	7.38 ±2.19***	**37.79 ±5.77**	12.60 ±0.34***
PDTB	51.97 ±0.48	6.73 ±0.94***	35.84 ±2.42	**55.02 ±0.38* **

(III) Pair Classification

	Test Set					
Training Set	All	AltLex	BECAUSE	CTB	PDTB	SemEval
All	**84.68 ±0.27**	**84.76 ±0.66**	**91.21 ±1.18**	**73.94 ±4.68**	83.28 ±0.36	94.71 ±0.23
AltLex	31.83 ±3.93***	80.57 ±2.48*	48.44 ±20.00**	20.06 ±7.14***	25.11 ±8.75***	57.72 ±14.52**
BECAUSE	36.40 ±0.64***	47.99 ±1.33***	90.01 ±1.95	23.58 ±1.52***	38.39 ±0.37***	25.23 ±2.02***
CTB	20.17 ±5.78***	19.16 ±15.64***	22.00 ±10.92***	73.29 ±6.14	7.02 ±6.06***	63.69 ±5.65***
PDTB	68.13 ±0.88***	40.34 ±1.52***	82.59 ±2.17***	26.74 ±2.42***	**83.70 ±0.34**	33.64 ±1.76***
SemEval	26.66 ±1.86***	37.07 ±6.58***	25.70 ±11.46***	50.63 ±1.74***	8.08 ±3.20***	**94.80 ±0.28**

Table 7. Mean and standard deviation of F1 score across different training and test set combinations.

1) Training on all datasets returned the best performance across all tasks by a large margin.
2) The generalized model trained on all datasets did not always return the best performance for each corpus

We propose UniCausal, a unified resource and benchmark for causal text mining.

- Our codes were designed to allow researchers to work on **some or all datasets and tasks**, while still comparing their performance fairly against us or others. Researchers **can easily include new datasets** too.
- We **provided evaluation metrics per dataset as an initial benchmark** for future researchers to compete against.
- Our **codes and processed data is available online**. Our trained **baseline model checkpoints are uploaded** to Huggingface Hub.
- We believe that a unified model that learns from diverse objectives and knowledge sources will be more adaptable and generalizable. We hope to see researchers build such models in the future using our repository.

Thank you.

- Link to our repository: https://github.com/tanfiona/UniCausal
- For questions/ feedback, feel free to contact us:

Fiona Anting Tan
Institute of Data Science
National University of Singapore, Singapore
tan.f@u.nus.edu