Event Causality Identification with Causal News Corpus - Shared Task 3, CASE 2022

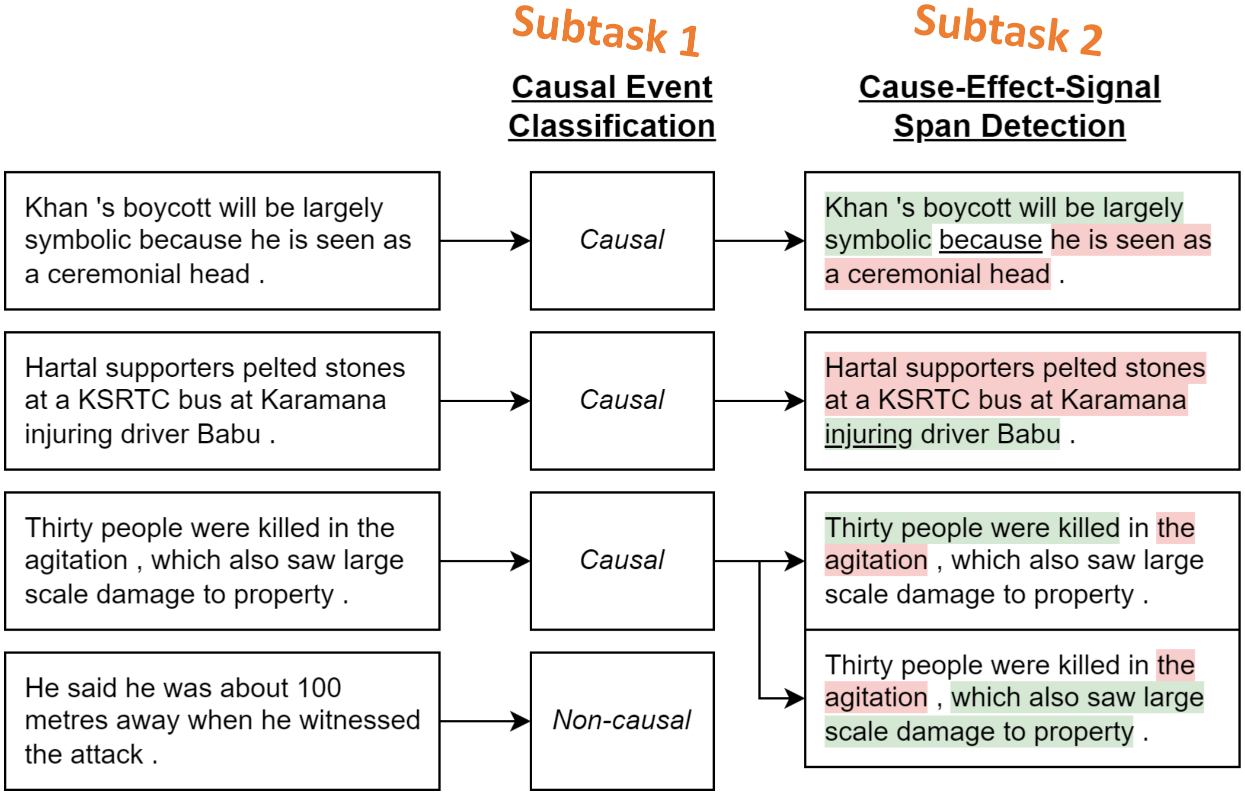
Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoglu, Tommaso Caselli, Onur Uca, Farhana Ferdousi Liza, Nelleke Oostdijk

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

TASK

01

Event Causality Identification Shared Task involved two subtasks related to Classification and Span Detection.



Subtask 1

Causal Event Classification

Subtask 2

Cause-Effect-Signal Span Detection

Khan 's boycott will be largely symbolic because he is seen as a ceremonial head .

Causal

Khan 's boycott will be largely symbolic because he is seen as a ceremonial head .

Hartal supporters pelted stones at a KSRTC bus at Karamana injuring driver Babu .

Causal

Hartal supporters pelted stones at a KSRTC bus at Karamana injuring driver Babu .

Thirty people were killed in the agitation , which also saw large scale damage to property .

Causal

Thirty people were killed in the agitation , which also saw large scale damage to property .

Thirty people were killed in the agitation , which also saw large scale damage to property .

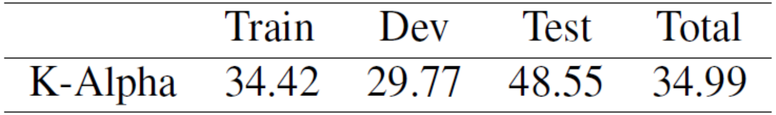
He said he was about 100 metres away when he witnessed the attack .

Non-causal

DATASET

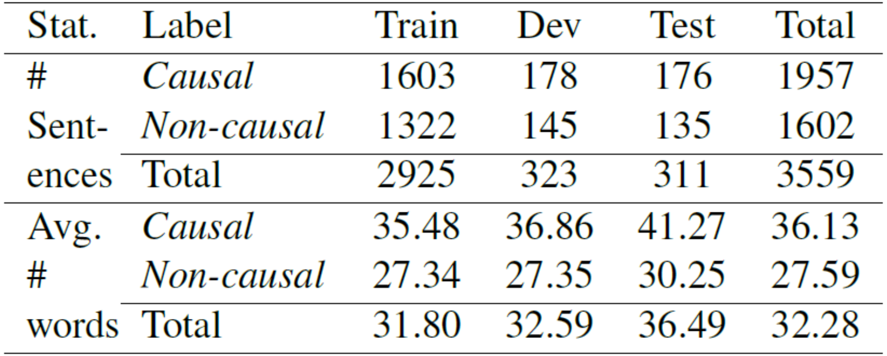
02

Subtask 1 worked directly on the Causal News Corpus (CNC) (Tan et al., 2022).



	Train	Dev	Test	Total
K-Alpha	34.42	29.77	48.55	34.99

Subtask 1 Inter-annotator Agreement Scores. Reported in percentages.



Stat.	Label	Train	Dev	Test	Total
# Sent-ences	*Causal*	1603	178	176	1957
	Non-causal	1322	145	135	1602
	Total	2925	323	311	3559
Avg. # words	*Causal*	35.48	36.86	41.27	36.13
	Non-causal	27.34	27.35	30.25	27.59
	Total	31.80	32.59	36.49	32.28

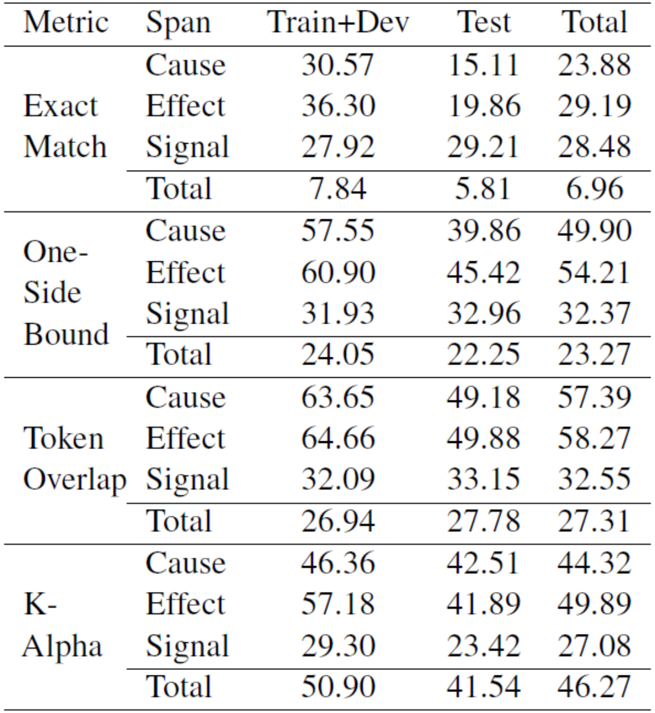
Subtask 1 Data Summary Statistics.

- Data Source:
  Causal News Corpus (CNC) (Tan et al., 2022)
  - 869 news documents
  - 3,559 English sentences
- A sentence is *Causal* if “one argument provides the reason, explanation or justification for the situation described by the other”(Webber et al., 2019) and contains at least a pair of events.

DATASET

03

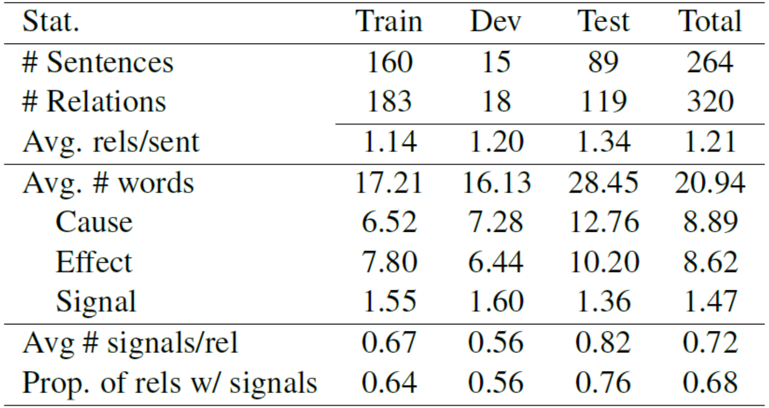
We added annotations for some *Causal* sentences from CNC with Cause, Effect and Signal spans for Subtask 2.



Metric	Span	Train+Dev	Test	Total
Exact Match	Cause	30.57	15.11	23.88
	Effect	36.30	19.86	29.19
	Signal	27.92	29.21	28.48
	Total	7.84	5.81	6.96
One-Side Bound	Cause	57.55	39.86	49.90
	Effect	60.90	45.42	54.21
	Signal	31.93	32.96	32.37
	Total	24.05	22.25	23.27
Token Overlap	Cause	63.65	49.18	57.39
	Effect	64.66	49.88	58.27
	Signal	32.09	33.15	32.55
	Total	26.94	27.78	27.31
K-Alpha	Cause	46.36	42.51	44.32
	Effect	57.18	41.89	49.89
	Signal	29.30	23.42	27.08
	Total	50.90	41.54	46.27

Subtask 2 Inter-annotator Agreement Scores. Reported in percentages.

- A **Cause** is a reason, explanation or justification that led to an **Effect**.
- **Signals** are words that help to identify the structure of the discourse.



Stat.	Train	Dev	Test	Total
# Sentences	160	15	89	264
# Relations	183	18	119	320
Avg. rels/sent	1.14	1.20	1.34	1.21
Avg. # words	17.21	16.13	28.45	20.94
Cause	6.52	7.28	12.76	8.89
Effect	7.80	6.44	10.20	8.62
Signal	1.55	1.60	1.36	1.47
Avg # signals/rel	0.67	0.56	0.82	0.72
Prop. of rels w/ signals	0.64	0.56	0.76	0.68

Subtask 2 Data Summary Statistics.

04

We provided multiple evaluation metrics, but model performance was eventually ranked by F1.

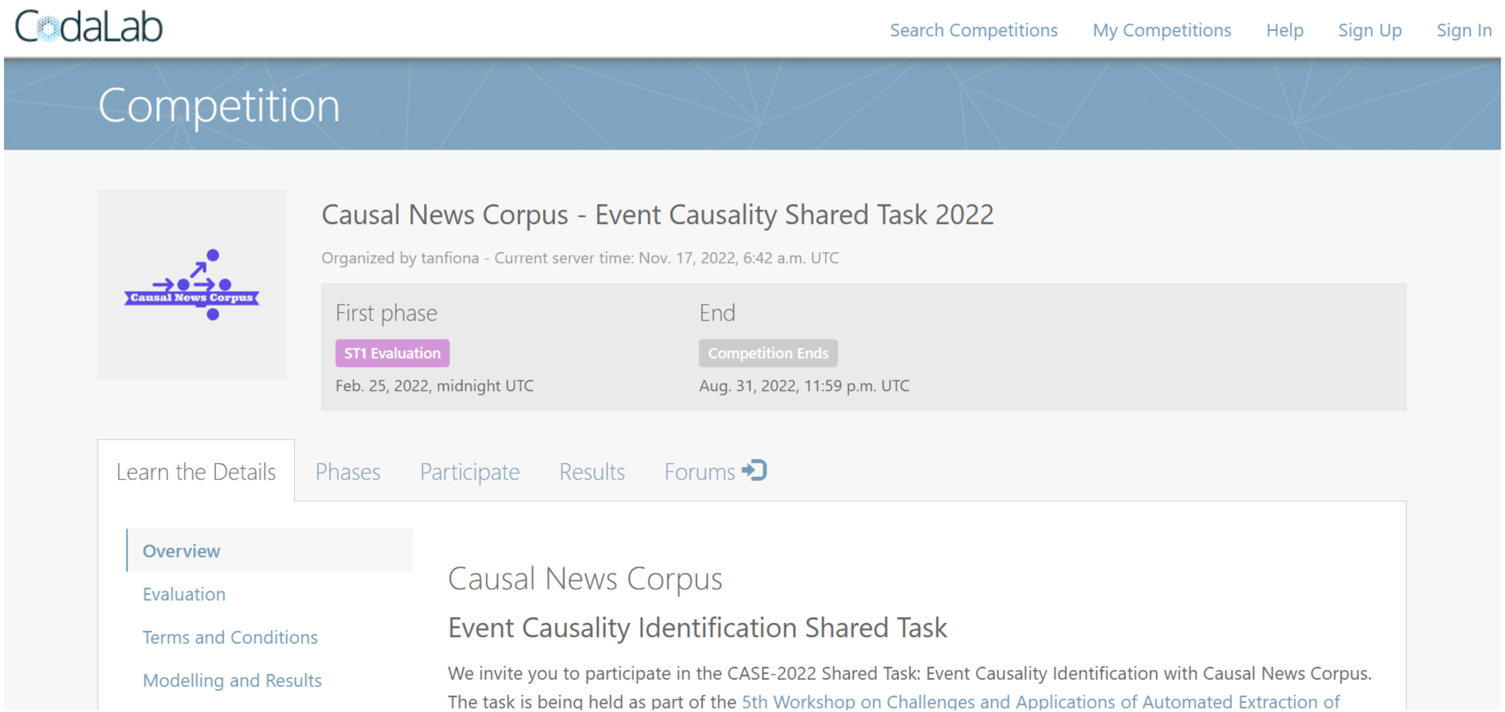
- The following evaluation metrics were provided:
  - Subtask 1: Accuracy, Binary Precision (P), Binary Recall (R), Binary F1 and Matthews Correlation Coefficient
  - Subtask 2: Macro P, R and F1 based on word labels
- Leader board was ranked by F1 for both tasks

- For Subtask 2, to handle predictions for examples with multiple causal relations:
  - If more predictions (p) are provided than true relations (n), we only consider the first n relations.
  - If fewer predictions (p) are provided than true relations (n), we assume the missing n-p relations have all “Other” tokens.
  - Once n=p, we calculate every combination of pairs of prediction and true relations and retain the combination that gives us the highest score.

COMPETITION

05

We used the Codalab website to host our competition.



Causal News Corpus - Event Causality Shared Task 2022

Organized by tanfiona - Current server time: Nov. 17, 2022, 6:42 a.m. UTC

First phase	End
ST1 Evaluation	Competition Ends
Feb. 25, 2022, midnight UTC	Aug. 31, 2022, 11:59 p.m. UTC

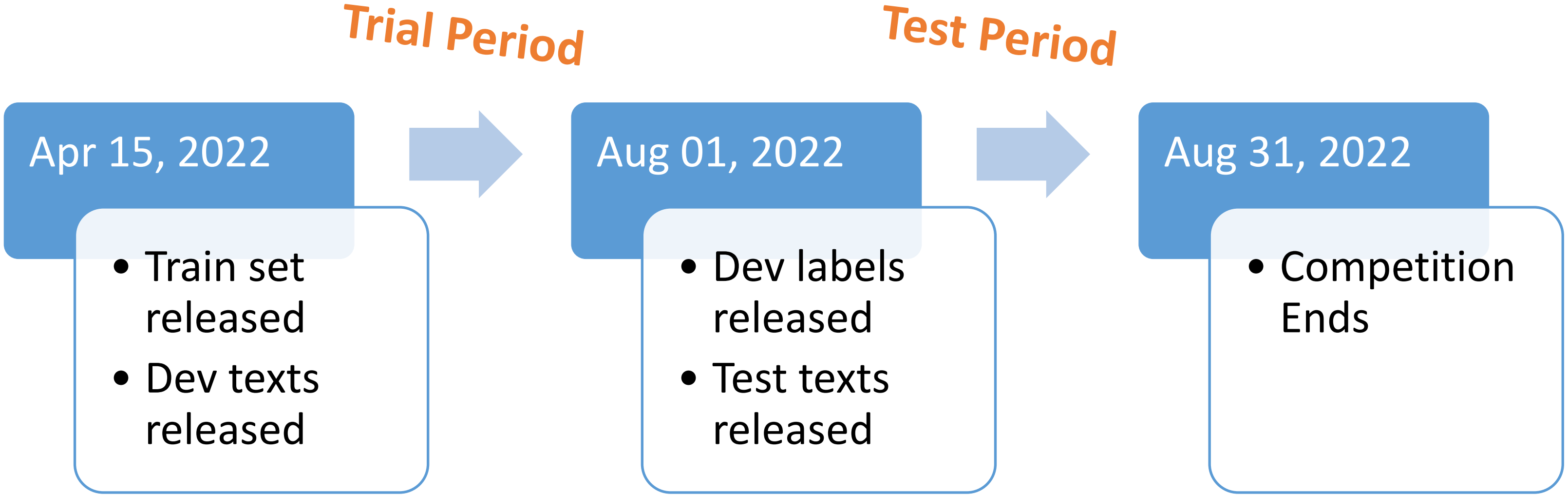
Causal News Corpus

Event Causality Identification Shared Task

We invite you to participate in the CASE-2022 Shared Task: Event Causality Identification with Causal News Corpus. The task is being held as part of the 5th Workshop on Challenges and Applications of Automated Extraction of

https://codalab.lisn.upsaclay.fr/competitions/2299#learn_the_details

06 Timeline



Trial Period

Test Period

Apr 15, 2022

- Train set released
- Dev texts released

Aug 01, 2022

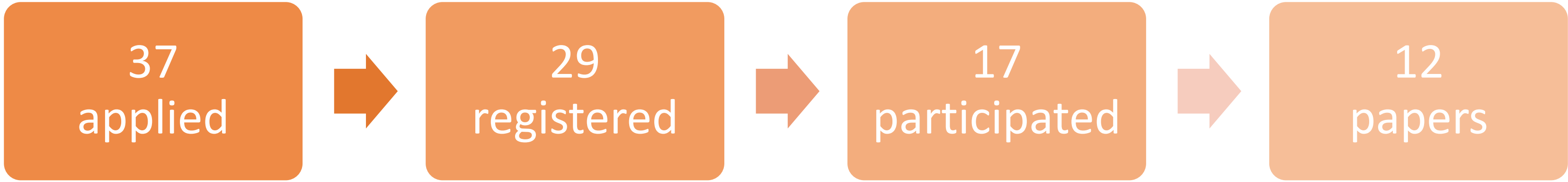
- Dev labels released
- Test texts released

Aug 31, 2022

- Competition Ends

Timeline of competition.

07 There were 17 active participants who made over 100 submissions on the test set.



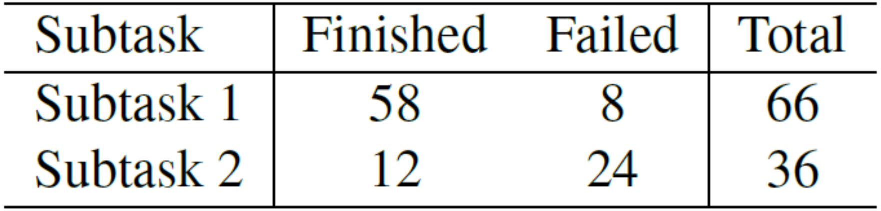
37 applied

29 registered

17 participated

12 papers

Number of teams per stage of competition.

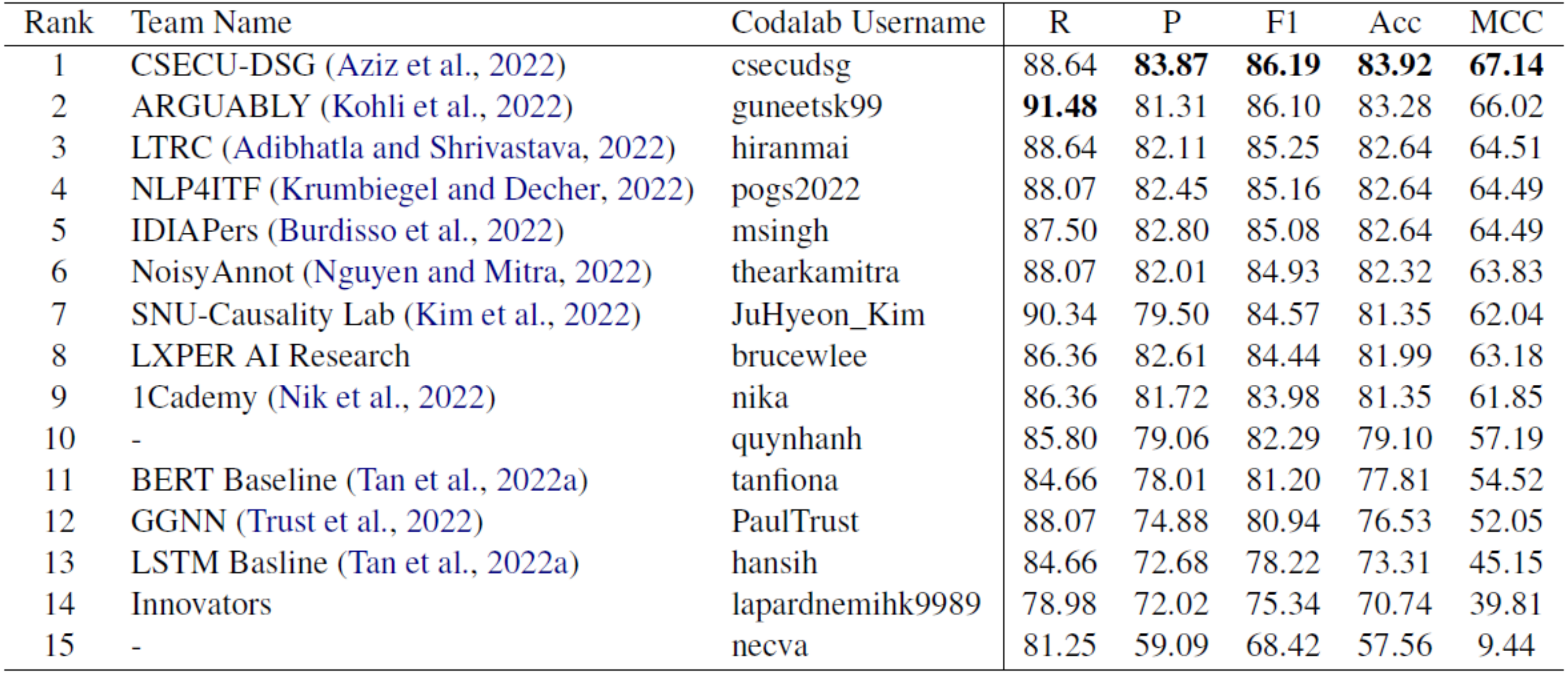


Subtask	Finished	Failed	Total
Subtask 1	58	8	66
Subtask 2	12	24	36

Number of submissions received for test set.

RESULTS

08 The best F1 score for Subtask 1 was 86.19%.



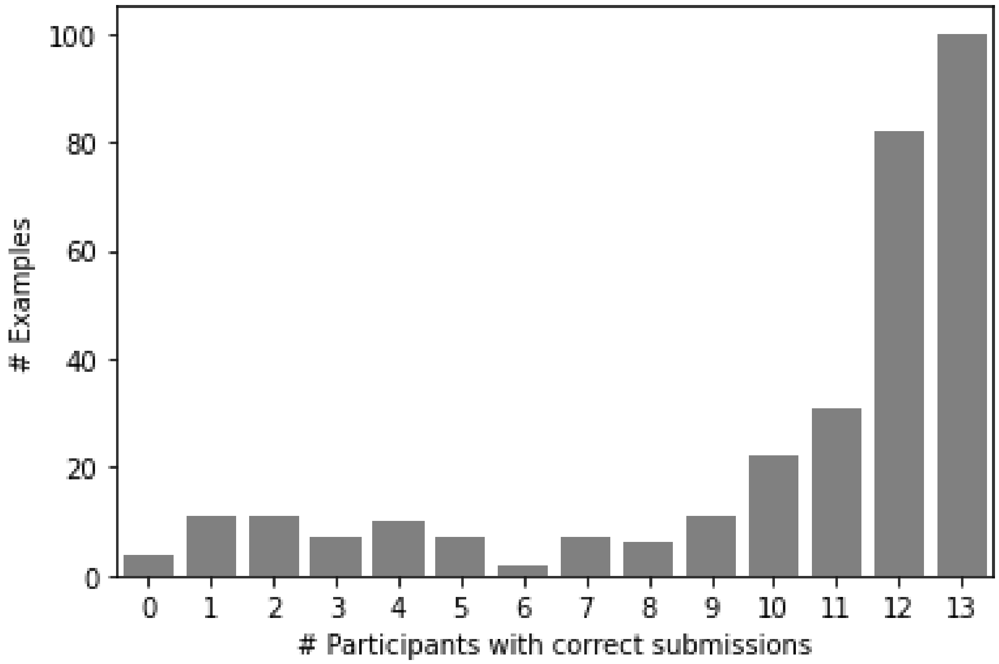
Rank	Team Name	Codalab Username	R	P	F1	Acc	MCC
1	CSECU-DSG (Aziz et al., 2022)	csecudsg	88.64	**83.87**	**86.19**	**83.92**	**67.14**
2	ARGUABLY (Kohli et al., 2022)	guneetsk99	**91.48**	81.31	86.10	83.28	66.02
3	LTRC (Adibhatla and Shrivastava, 2022)	hiranmai	88.64	82.11	85.25	82.64	64.51
4	NLP4ITF (Krumbiegel and Decher, 2022)	pogs2022	88.07	82.45	85.16	82.64	64.49
5	IDIAPers (Burdisso et al., 2022)	msingh	87.50	82.80	85.08	82.64	64.49
6	NoisyAnnot (Nguyen and Mitra, 2022)	thearkamitra	88.07	82.01	84.93	82.32	63.83
7	SNU-Causality Lab (Kim et al., 2022)	JuHyeon_Kim	90.34	79.50	84.57	81.35	62.04
8	LXPER AI Research	brucewlee	86.36	82.61	84.44	81.99	63.18
9	1Cademy (Nik et al., 2022)	nika	86.36	81.72	83.98	81.35	61.85
10	-	quynhanh	85.80	79.06	82.29	79.10	57.19
11	BERT Baseline (Tan et al., 2022a)	tanfiona	84.66	78.01	81.20	77.81	54.52
12	GGNN (Trust et al., 2022)	PaulTrust	88.07	74.88	80.94	76.53	52.05
13	LSTM Basline (Tan et al., 2022a)	hansih	84.66	72.68	78.22	73.31	45.15
14	Innovators	lapardnemihk9989	78.98	72.02	75.34	70.74	39.81
15	-	necva	81.25	59.09	68.42	57.56	9.44

Subtask 1 Leaderboard.

ERROR ANALYSIS

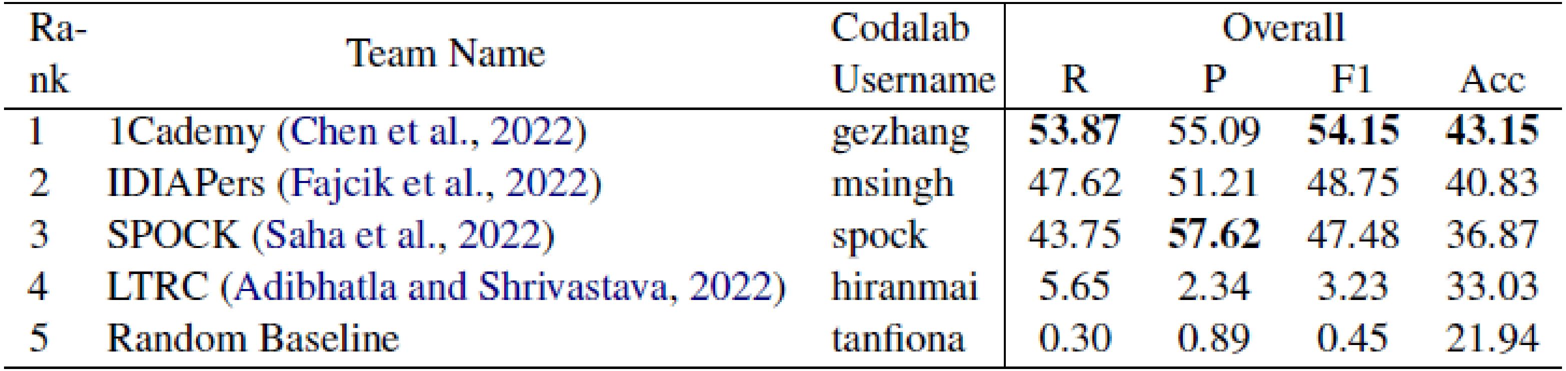
09

Many examples (100/311) in the test set could be predicted correctly by all participants.



RESULTS

10 The best F1 score for Subtask 2 was 54.15%.



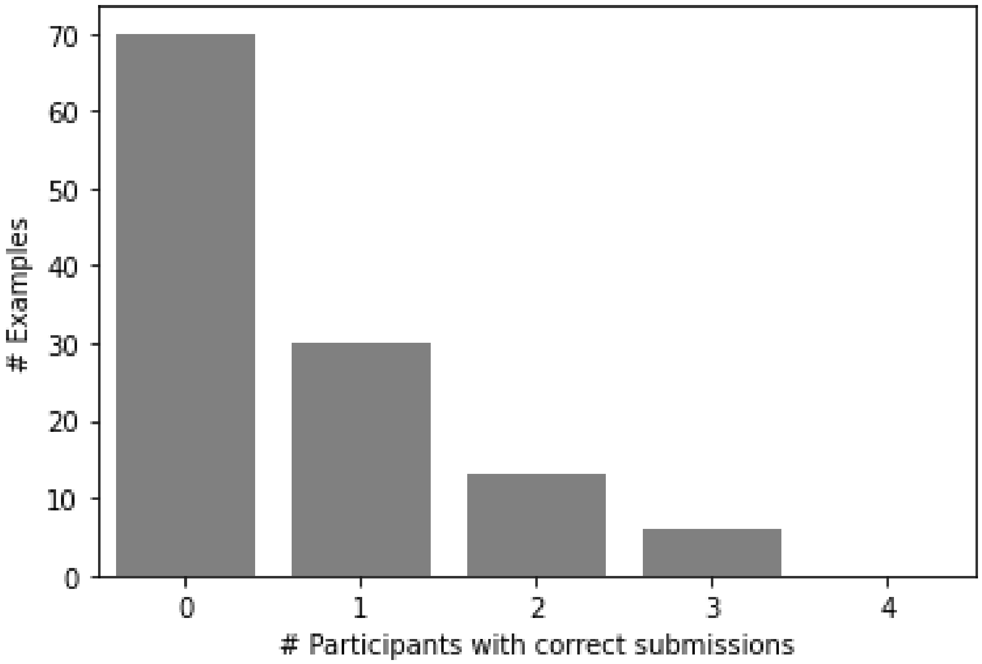
Ra-nk	Team Name	Codalab Username	Overall			
			R	P	F1	Acc
1	1Cademy (Chen et al., 2022)	gezhang	**53.87**	55.09	**54.15**	**43.15**
2	IDIAPers (Fajcik et al., 2022)	msingh	47.62	51.21	48.75	40.83
3	SPOCK (Saha et al., 2022)	spock	43.75	**57.62**	47.48	36.87
4	LTRC (Adibhatla and Shrivastava, 2022)	hiranmai	5.65	2.34	3.23	33.03
5	Random Baseline	tanfiona	0.30	0.89	0.45	21.94

Subtask 2 Leaderboard.

ERROR ANALYSIS

11

Most examples were predicted wrongly by all participants.



Conclusion & Future Work

- Two subtasks:
  1) Causal Event Classification, and
  2) Cause-Effect-Signal Span Detection.
- Each subtask attracted predictions from models that beat our baselines.

- Next iteration:
  - More data for Subtask 2!

Thank you.

Fiona Anting Tan
tan.f@u.nus.edu
https://github.com/tanfiona/CausalNewsCorpus