# 1Cademy @ Causal News Corpus 2022: Leveraging Self-Training in Causality Classification of Socio-Political Event Data

**Adam Nik[2 4], Ge Zhang[1 2 3], Xingran Chen[3], Mingyu Li[2 3], Jie Fu[1]**

[1] Beijing Academy of Artificial Intelligence, China

[2] 1Cademy Community, USA

[3] University of Michigan Ann Arbor, USA

[4] Carleton College, USA

## Abstract

This paper details our participation in the Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) workshop @ EMNLP 2022, where we take part in Subtask 1 of Shared Task 3[1]. We approach the given task of event causality detection by proposing a self-training pipeline that follows a teacher-student classifier method. More specifically, we initially train a teacher model on the true, original task data, and use that teacher model to self- label data to be used in the training of a separate student model for the final task prediction. We test how restricting either the number of positive or negative self-labeled examples in the self-training process affects classification performance. Our final results show that using self-training produces a comprehensive performance improvement across all models and self-labeled training sets tested within the task of event causality sequence classification. On top of that, we find that self-training performance did not diminish even when restricting either positive/negative examples used in training.

## Research Task:

Task 1 of the CASE workshop @ EMNLP 2022 works to identify and classify event causality in socio-political event (SPE) data, with subtask 1 being a binary classification of causality. In other words, participants are tasked with answering: Does an event sentence contain a cause-effect relationship?

## Causal News Corpus Data

The CNC dataset[2] is a corpus of 3,559 event sentences from protest event news labeled on whether a given sentence contains causal relations or not. The data of the CNC comes from two workshops focused on mining socio-political data: Automated Extraction of Socio-political Events from News (AESPEN)[3] in 2020 and the CASE 2021 workshop @ ACL-IJCNLP[4]. For purposes of subtask 1, the data was split into a training set of 2925 examples, a development set of 3,23 examples, and a final test set of 311 examples that was used as an evaluation benchmark for the competition.

## Self-Training Methodology

We follow a similar teacher-student pipeline as Yalniz et al.[5] that includes using a teacher model to generate a new labeled dataset $D'$ from the original dataset $D$ and then training a new student model on the new labeled dataset $D'$ and then on the original dataset $D$. We used the training split provided of 2925 CNC samples[2] as the original dataset $D$, and fine-tune a BERT model[6] for sequence classification, which served as our teacher model.

We can improve the performance of **classification models** by simply augmenting the training data with **self-labeled examples**.

Additionally, performance improvements from **self-training do not diminish** when either positive or negative self-labeled examples are **restricted**.

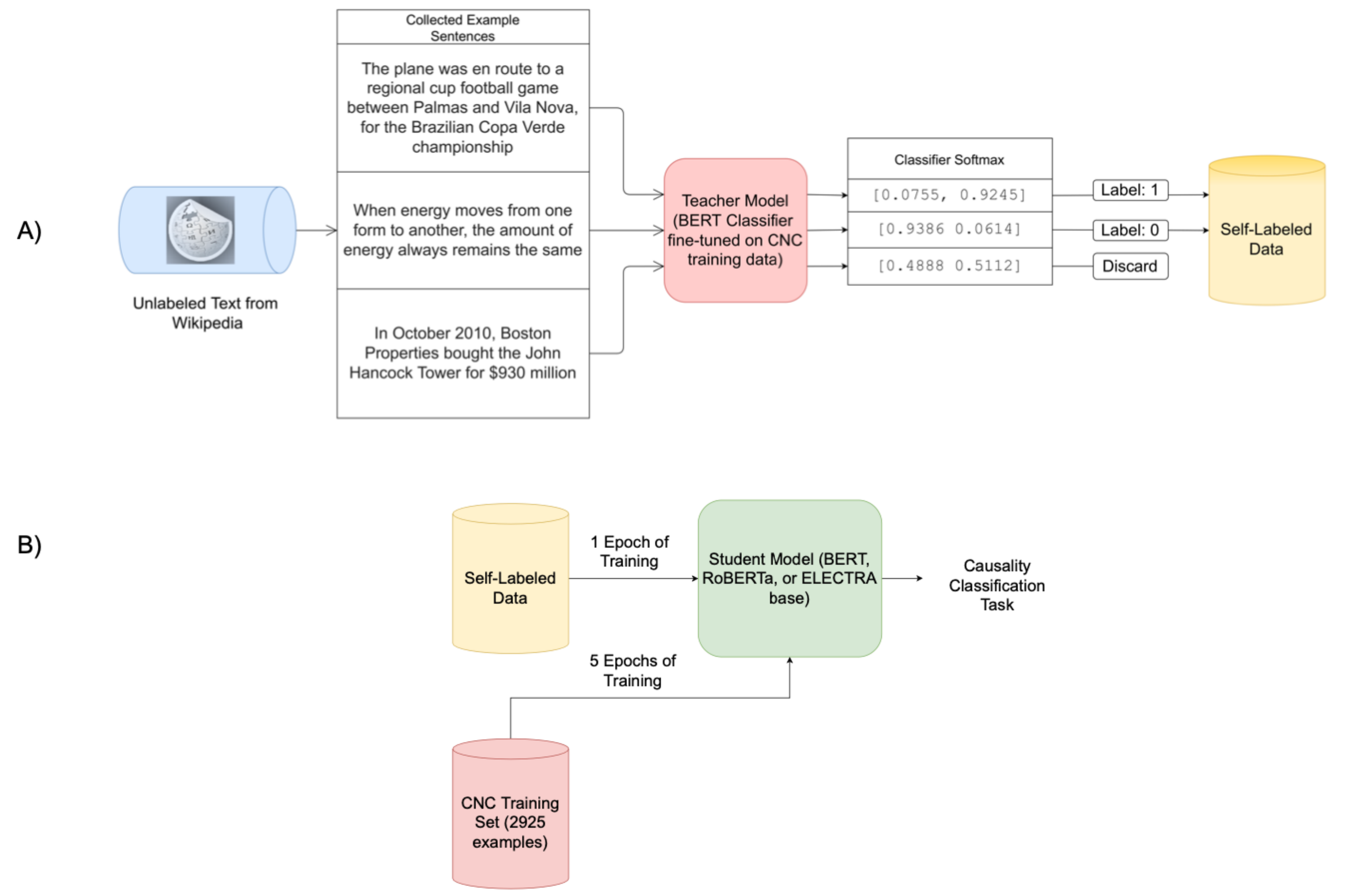


Figure 1: A) Self-training pipeline with Teacher Model. B) We use the self-labeled examples as part of the training when training in Student Models for the task of causality classification

## Experiment Set-up

In our experimentation set-up, we tested all three backbone models—BERT[6], RoBERTa[7] , and ELECTRA[8]—with both the self-training pipeline and a simple fine-tuning process that only used the provided CNC training set which served as our baseline. In the baseline experiments, the classifiers were trained solely on five epochs of the CNC training data.

**Baseline Training vs. Self-Training Results**

| | | | Accuracy | F1 | Recall | Precision | MCC |
|---|---|---|---|---|---|---|---|
| Baseline Training (simple fine-tuning, no self-training) | BERT | | 0.8204 | 0.8394 | 0.8516 | 0.8276 | 0.6363 |
| | RoBERTa | | 0.8390 | 0.8543 | 0.8561 | 0.8525 | 0.6745 |
| | Google ELECTRA Discriminator | | 0.8365 | 0.8535 | 0.8640 | 0.8432 | 0.6689 |
| | | Ratio of Positive to Negative Self-Labeled Examples used in training | Accuracy | F1 | Recall | Precision | MCC |
| Self-Training | BERT | 1:3 | 0.8380 | 0.8531 | 0.8539 | 0.8525 | 0.6726 |
| | | 1:1 | 0.8225 | 0.8377 | 0.8315 | 0.8468 | 0.6425 |
| | | 3:1 | 0.8380 | 0.8526 | 0.8502 | 0.8552 | 0.6728 |
| | RoBERTa | 1:3 | 0.8576 | 0.8715 | **0.8764** | 0.8671 | 0.7123 |
| | | 1:1 | **0.8586** | 0.8711 | 0.8670 | **0.8755** | **0.7149** |
| | | 3:1 | **0.8586** | **0.8719** | 0.8727 | 0.8711 | 0.7142 |
| | Google ELECTRA Discriminator | 1:3 | 0.8400 | 0.8579 | **0.8764** | 0.8415 | 0.6760 |
| | | 1:1 | 0.8524 | 0.8665 | 0.8689 | 0.8641 | 0.7016 |
| | | 3:1 | 0.8421 | 0.8580 | 0.8652 | 0.8510 | 0.6806 |

Table 1: Results of the evaluating the CNC development set on both simple fine-tuning with only CNC training data (top) and fine-tuning classifiers on training sets of self-labeled data in addition to CNC training data (bottom). **Bold** indicates highest performance across all splits and model types, underline indicates the highest performance of the specific model type.

## Results and Discussion

From the table, we can see that every self-training setup outperformed the baseline classifier in terms of accuracy, with an average accuracy improvement of 1.33% across all models and polarity splits. Furthermore, for all but one self-training set-up, there was an improvement of the F1 score from the baseline, with an average improvement of 0.011.

Our results show that training a classifier on self-labeled data using a teacher-student approach comprehensively improves task performance. Furthermore, we find that performance improvement from self-training did not differ significantly between self-labeled training sets with varying levels of example polarity. This indicates that the model is capable of reaping the full benefits of self-training despite having limited access to positive or negative samples.

## References

1. Fiona Anting Tan, Ali Hürriyetog̀lu, Tommaso Caselli, Nelleke Oostdijk, Hansi Hettiarachchi, Tadashi Nomoto, Onur Uca, and Farhana Ferdousi Liza. 2022a. Event causality identification with causal news corpus - shared task 3, CASE 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, Online. Association for Computational Linguistics.
2. Fiona Anting Tan, Ali Hürriyetog̀lu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi He tiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022b. The causal news corpus: Annotating causal relations in event sentences from news. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2298– 2310, Marseille, France. European Language Resources Association.
3. Ali Hürriyetog̀lu, Vanni Zavarella, Hristo Tanev, Erdem Yörük, Ali Safaya, and Osman Mutlu. 2020. Automated extraction of socio-political events from news (AESPEN): Workshop and shared task report. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 1– 6, Marseille, France. European Language Resources Association (ELRA).
4. Ali Hürriyetog̀lu, Hristo Tanev, Vanni Zavarella, Jakub Piskorski, Reyyan Yeniterzi, Osman Mutlu, Deniz Yuret, and Aline Villavicencio. 2021. Challenges and applications of automated extraction of socio- political events from text (CASE 2021): Workshop and shared task report. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 1–9, Online. Association for Computational Linguistics.
5. I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi- supervised learning for image classification.
6. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
7. 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.
8. Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.