Team "NoConflict" at CASE 2021 Task 1: Pretraining for Sentence-Level Protest Event Detection

Tiancheng Hu **Niklas Stoehr**
ETH Zurich, Switzerland
tianhu@student.ethz.ch niklas.stoehr@inf.ethz.ch

Abstract

An ever-increasing amount of text, in the form of social media posts and news articles, gives rise to new challenges and opportunities for the automatic extraction of socio-political events. In this paper, we present our submission[1] to the Shared Tasks on Socio-Political and Crisis Events Detection, Task 1, Multilingual Protest News Detection, Subtask 2, Event Sentence Classification, of CASE @ ACL-IJCNLP 2021. In our submission, we utilize the RoBERTa model with additional pretraining, and achieve the best F1 score of 0.8532 in event sentence classification in English and the second-best F1 score of 0.8700 in Portuguese via simple translation. We analyze the failure cases of our model. We also conduct an ablation study to show the effect of choosing the right pretrained language model, adding additional training data and data augmentation.

1 Introduction

With the growing volume of online news from both traditional news media and social media, large amounts of texts are being created every day. These text data contain information about events happening around the world. For social science and policy making, the event information in these texts can be extremely valuable. Due to the sheer volume of data available, there is a strong demand for tools to automatically extract and analyze socio-political events. Automatic event extraction enables governments, non-governmental organizations and society as a whole to take more timely, proportional and appropriate actions in changing circumstances.

Event sentence classification is an important step in the event extraction pipeline (Hürriyetoğlu et al., 2019a). In this work, we present our submission to the CASE 2021 Shared Task, hosted jointly with the workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) @ ACL-IJCNLP 2021 (Hürriyetoğlu et al., 2021). The shared task consists of two main tasks: Multilingual Protest News Detection and Fine-Grained Classification of Socio-Political Events. In the first shared task, there are four subtasks: event document classification, event sentence classification, event sentence coreference resolution and event extraction. In this paper, we focus on Task 1, Subtask 2, namely event sentence classification. For a detailed description of the shared task, please refer to Hürriyetoğlu et al. (2021). Prior iterations of the workshop can be found in Hürriyetoğlu et al. (2019b, 2020).

Within this subtask, we further narrow down our scope by focusing on the English event descriptions only. We train a classifier to solve the binary classification problem to identify whether a sentence contains a protest event, as defined in Hürriyetoğlu et al. (2021). Given the huge success of pretrained language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019) and ELECTRA (Clark et al., 2020), we adopt RoBERTa as the backbone of our model. Inspired by the good result achieved through additional pretraining (Gururangan et al., 2020), we harness the POLUSA dataset (Gebhard and Hamborg, 2020) of political news articles to second-pretrain our model with a masked language modeling (MLM) objective. Further, we conduct a series of ablation studies to justify our design choices. We first run experiments to choose a suitable base model. Then, we conduct experiments on using additional training data from other subtasks. Given the limited amount of training data available, we experiment with data augmentation techniques, including back translation, embedding augmentation, and checklist augmentation (Ribeiro et al., 2020). The rest of this paper is organized

[1] Code available at https://github.com/pitehu/CASE_2021

Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) , pages 152–160

as follows: Section 2 describes the dataset of the subtask. Section 3 discusses the method of our best-performing submission. Section 4 presents quantitative results achieved by our model as well as a failure case analysis. In Section 5, we present additional experiments as part of an ablation study. In Section 6, we discuss observations of the dataset and models trained on this dataset from the perspective of named entities before concluding the paper in Section 7.

2 Dataset

We are provided with a dataset of labeled sentences which was introduced in Hürriyetoğlu et al. (2021). Each sentence has a binary label indicating if the sentence contains a protest event. While the dataset comprises sentences in English, Spanish and Portuguese, we solely focus on English sentences. The English version of this dataset contains 22,825 sentences, out of which 18,602 (81.50%) have label 0 and 4223 (18.50%) have label 1. Since no official train-validation split is provided, we divide the dataset into a training set (80%) and a validation set (20%).

3 Proposed Method

We utilize the RoBERTa base model (Liu et al., 2019) as the backbone of our model. Throughout this work, we refer to the pretrained RoBERTa model (Liu et al., 2019) as "RoBERTa default". We use the term language model to refer to *Transformer*-based (Vaswani et al., 2017) cloze language models.

Second Pretraining We start by conducting an additional round of pretraining of RoBERTa, initialized with the already pretrained weight, following Gururangan et al. (2020). To this end, we pretrain on the POLUSA dataset (Gebhard and Hamborg, 2020) in an MLM setting with a masking probability of 0.15. We denote this pretraining step as *Second Pretraining*. Intuitively, language models are usually trained on large and diverse datasets of different domains. Thus, their language modeling capacity may not be optimal in specific domains such as protest event classification.

POLUSA Dataset The POLUSA dataset (Gebhard and Hamborg, 2020) is a dataset containing political news covering policy topics published between January 2017 and August 2019. It contains about 0.9M news articles from 18 outlets representing the political spectrum.

Finetuning Once the *second pretraining* is completed, we feed the [CLS] embedding of the last hidden layer to a fully-connected layer, which serves as the classification head. The [CLS] embedding encodes information of the whole sentence.

4 Results

We conduct *second pretraining* for only 42,000 steps, with a batch size of 16 and a maximum sequence length of 256, due to time and resource constraints. For the downstream task, we train for 25 epochs and take the best epoch based on validation F1 score.

4.1 Quantitative Results

Second Pretraining In this section, we discuss the effect of the *second pretraining*. We take a checkpoint every 4000 steps and finetune for the event sentence classification task, and report the best F1 score and the MLM loss during the pretraining in Figure 1. Additionally, we manually select 10 representative sentences from the Subtask 2 dataset and measure the average change of their representations in embedding space during *second pretraining*. To this end, we compute the Euclidean distance between the embedding of a sentence yield by the RoBERTa default model and by our model during *second pretraining* at every checkpoint.

Finetuning Our model with the *second pretraining* strategy achieves a 0.8395 F1 score on the validation set of this subtask. On the evaluation server, we achieve the best performance among all submissions of the shared task with an F1 score of 0.8532 on the testing set. Since our focus is on the English version of the event sentence classification task, we translate the event sentences of other languages into English using Argos Translate (Finlay, 2021). This simple method achieves the second best F1 score of 0.8670 in Portuguese.

Failure Cases Investigating cases in which our model fails to classify sentences correctly offers helpful insights. The model's failure cases broadly fall into the following categories:

Semantic Error: the model makes a clear semantic error. An example is provided in Table 1. Sentence 1 does not contain a protest event but the model predicts one.

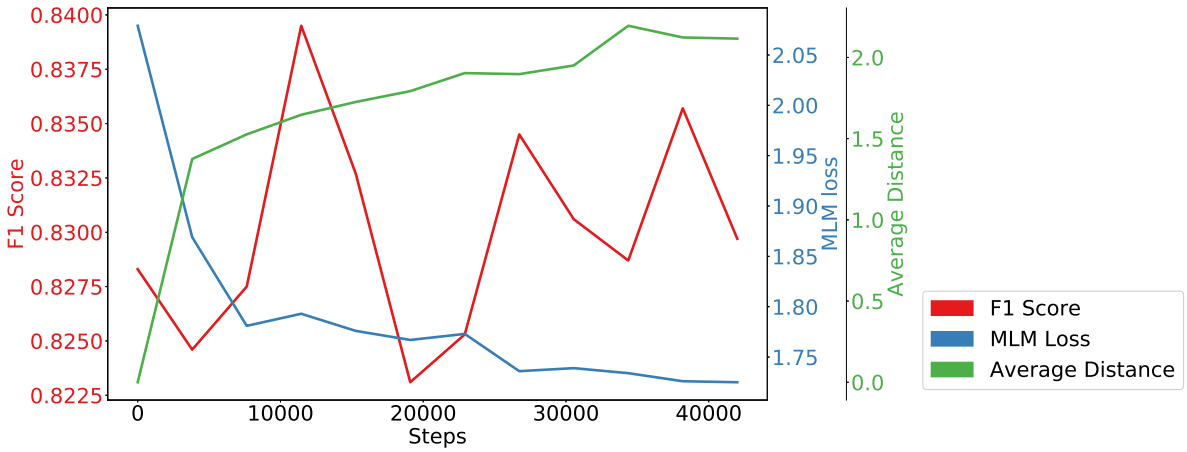

Figure 1: F1 Score, MLM loss, and embedding shift at different steps during the *second pretraining* phase. We take a checkpoint at every 4000 steps. The figure displays the MLM loss from the pretraining objective for each checkpoint as well as the validation F1 score from finetuning for the sentence classification task with this checkpoint. Additionally, we track how much the embeddings change by manually selecting 10 sentences. We measure the Euclidean distance between their vector representations from the RoBERTa default model and our model at each checkpoint. Best viewed in color.

Rule Error: this happens when the sentence could be seen as a protest event sentence in commonsense but is not considered one according to the annotation manual (Hürriyetoğlu et al., 2021). In Sentence 2 in Table 1, there is no indication that the event has happened already or is ongoing. Therefore, based on the annotation manual, it should be classified as negative while the model gives a positive prediction.

Uncertain Reference: The event that a sentence is referring to is ambiguous. We show two examples in Table 1 in sentences 3 and 4. "The act" and "this" refer to an event that we do not have knowledge of without context. In this subtask, we do not have access to any context and thus the labels for these two sentences are uncertain. However, they have opposite labels in the ground truth. This may pose difficulty for model training.

Indirect Mention: There are cases of label inconsistency when an event is indirectly mentioned. In Table 1, sentences 5 and 6 should both receive a positive label as they both pertain to a clear conflict event, but only sentence 6 has a positive label.

5 Ablation Study

In this section, we explore different base models to finetune on, using additional training data from other subtasks and data augmentation techniques. All results reported in this section are on the validation set, without *second pretraining*.

5.1 Base Model

First, we compare the performance of different base models. We consider BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019) and ELECTRA(Clark et al., 2020) as they represent some of the best-performing language models. Due to resource limitations, we only consider the base version of these models. We follow the same procedure as introduced in Section 3 but without the *second pretraining* step. We present the results in Table 2. We find that RoBERTa achieves the best results while BERT, XLNet and ELECTRA perform similarly or worse.

5.2 Additional Training Data

In this subsection, we explore adding data from other languages of Subtask 2 as well as from other subtasks. When we add data not originally in English, we translate the sentences into English using Argos Translate (Finlay, 2021). We present the result in Table 4. For example, "Sub1 ES&PT+" means Spanish and Portuguese data from Subtask 1 with positive labels. While some settings result in better performance, when we use them in conjunction with *second pretraining*, the performance gain disappears. Thus, we do not include any additional training data in training our model for final submission.

#	Sentence	P	L
	Semantic Error		
1	___ 9:05 a.m. Democratic presidential candidate Bernie Sanders is disavowing remarks made by a campaign surrogate who said voters shouldn't "continue to elect corporate Democratic whores" during a large New York City rally.	1	0
	Rule Error		
2	On Tuesday, a group of aviation staff called for a protest at Hong Kong airport on Friday to condemn the government and police for "ignoring the random attacks on citizens in Yuen Long".	1	0
	Uncertain Reference		
3	The act was captured by CCTV cameras and witnesses using smartphones.	1	0
4	"This has happened across the state.	0	1
	Indirect Mention		
5	He did not give details, but a local independent daily, O Pais, said six people were injured in the attack in Ancuabe in Mozambique's northern Cabo Delgado province.	1	0
6	Spokesman Keith Khoza said they had decided to March to Prime Media because the cartoon had raised various concerns.	0	1

Table 1: Example failure cases. We divide the failure cases into four categories and give example sentences of each category. "P" refers to the model's prediction and "L" refers to the ground truth label.

Base Model	F1 Score
BERT	0.8117
RoBERTa	**0.8283**
XLNet	0.8097
ELECTRA	0.8113

Table 2: Effect of Base Models. We keep all other settings fixed while changing the base models and conduct finetuning on event sentence classification.

Augmentation Methods	F1 Score
None	0.8283
Back Translation	0.8206
Embedding + Checklist	**0.8294**
Paraphrase	0.8026

Table 3: Effect of Data Augmentation. We train the event sentence classification model with augmented data from the data augmentation methods of Subtask 2 data, in addition to the original training data.

Multilingual Data from Subtask 2 In Experiment 2, we add Subtask 2 data from Spanish and Portuguese. We show the result in Experiment 2. The result nearly does not change.

Data from Subtask 3 and 4 In Experiment 3, we add data from Subtask 3 and 4. Subtask 3 is a event coreference resolution task and Subtask 4 is a event trigger detection task, both with data from protest events. As both are downstream tasks of Subtask 2 as shown in (Hürriyetoğlu et al., 2021), we assume that all sentences from the two subtasks contain event sentences and thus may help our Subtask 2 model. Upon manual inspection, there are some overlaps between Subtask 2 and Subtask 3 and 4 data but many new training samples exist. We see small gains compared to Experiments 1 and 2.

Combine Data from Subtask 2, 3 and 4 In Experiment 4, we combine the training data from Experiment 2 and 3, namely Subtask 2 data from all three languages and Subtask 3 data from English only. We see that the F1 score increases from 0.8303 to 0.8363. In Experiment 5, we also include Spanish and Portuguese Subtask 3 and 4 data. The F1 score is nearly the same as Experiment 4.

Negative Samples from Subtask 1 In Experiment 6, we add negative samples from the data of Subtask 1, in addition to the data from Experiment 5. Subtask 1 is a document classification task, in which "positive" indicates that the document contains a protest event. According to Hürriyetoğlu et al. (2020), a positive document contains protest event(s) but it does not imply that all sentences in that document should be labeled positive. On the other hand, any negative document is certain to contain no protest event. Therefore, we experiment with adding the negative documents first. The F1 score drops from 0.8362 to 0.8275. This is even worse than the baseline of considering only the

#	Sub1				Sub2		Sub3+4		F1
	EN+	EN-	ES&PT+	ES&PT-	EN	ES&PT	EN	ES&PT	
1					✓				0.8283
2					✓	✓			0.8282
3					✓		✓		0.8303
4					✓	✓	✓		**0.8363**
5					✓	✓	✓	✓	0.8362
6		✓			✓	✓	✓	✓	0.8275
6		✓		✓	✓	✓	✓	✓	0.8254
7	✓				✓	✓	✓	✓	0.7646
8	✓		✓		✓	✓	✓	✓	0.7439

Table 4: Effect of Training Data. In this table, we show the impact of having different combinations of training data from different subtasks. EN, ES and PT mean the English, Spanish, and Portuguese versions of the training data from a specific subtask, respectively. In Subtask 1, we consider the positive class and negative class separately. "+" indicates data from the positive class while "-" indicates data from the negative class.

English Subtask 2 data (Experiment 1). In Experiment 7, we add the translated negative samples from Spanish and Portuguese. The resulting F1 score further drops to 0.8254.

Positive Samples from Subtask 1 In Experiment 7, we add positive samples from the English version of Subtask 1, assuming that a positive document implies that every sentence in the document has a positive label. In Experiment 8, we add positive samples from the Spanish and Portuguese versions of Subtask 1. As we suspected, this assumption does not hold and the F1 scores drop significantly, to well below 0.8 in both cases.

5.3 Effect of Data Augmentation

In addition to adding more training data directly, we also consider data augmentation methods: back translation, checklist augmentation, embedding augmentation and paraphrasing. We point the reader to Section A.2 in the appendix for a description of these methods and example sentences generated with these augmentation methods. Some augmentation methods result in better performance. When combined with *second pretraining*, however, the performance gain disappears. Thus, we do not include any augmented data in training our model for final submission.

Results We show the result of the models trained with data augmentations in Table 3. We notice a drop in performance in back translation. This may be due to the subtle differences between translated sentences and task sentences native in English, similar to what we discussed in Section 5.2. We find a small improvement in embedding and checklist augmentations. We believe performing these two augmentations makes the model more robust to changes in contextual information. Paraphrasing results in a large drop in the model performance from 0.8283 without augmentation to just above 0.8 in F1 score. After inspecting the paraphrased sentences, we find that the paraphrasing model changes the input sentences very dramatically. In some cases, pieces of information that do not exist in the source text are even created.

Training Data	Initialization	F1 Score
No NE	RoBERTa	0.8210
No NE	second-pretrain	0.8277
Only NE	RoBERTa	0.3959
Only NE	second-pretrain	0.4190
Random	None	0.1896
All	RoBERTa	0.8283
All	second-pretrain	**0.8395**

Table 5: Effect of named entities. We finetune models with data without NEs and data with only NEs, with both RoBERTa default and RoBERTa *second pretraining*. We modify the validation data accordingly. The result shown is on the validation set. We include a random guessing baseline model for comparison. We also include the model performances with no modification to the data for reference.

6 Effect of Named Entities

In this section, we analyse the effect of named entities (NE) on the results. We train models with two modifications of the Subtask 2 data: 1. we remove all named entities in all sentences; 2. we

remove all text tokens except for named entities in each sentence. For each data modification setting, we train two models, one model initialized with the RoBERTa default weight, one initialized with the second-pretrained weight as mentioned in Section 3. For comparison, we include a random guessing baseline model. It draws label from the same distribution of the ground truth labels in the training set, without considering the sentences at all. We report the average F1 score of 100 such random assignments. We also include the result of the model trained with the original Subtask 2 training data using RoBERTa default weight and second-pretrain weight for reference. The result is shown in Table 5. We notice that without NEs, the model performs worse than the model trained with full data, in both RoBERTa default weight case and *second pretraining* case, suggesting that NEs contribute to the model's ability to correctly classify protest sentences. We also see that by only relying on NEs, the model is able to achieve an F1 score of around 0.4, more than double that of the random baseline, further suggesting that in this dataset, there are statistics about NEs that the model may utilize to make its decision, in addition to capturing linguistic clues. For example, due to the situation in Hong Kong in recent years, any sentence related to Hong Kong may have an above-average likelihood of containing an event. Additionally, we notice better performance in both the No NE setting and the Only NE setting when we finetune models with second pretrained weight. This emphasizes the importance of *second pretraining* in our approach.

7 Conclusion

In this paper, we present our submission to task 1, subtask 2 at CASE @ ACL-IJCNLP 2021. Our model is based on RoBERTa with a *second pretraining* step done on the POLUSA dataset. We inspect the failure cases of our model on the validation set and provide some explanations. To justify our design choices, we conduct an ablation study. Overall, we achieve the highest F1 score in the English version of this subtask and the second highest F1 score in Portuguese on the evaluation server. In future work, we plan to incorporate knowledge from the annotation manual into the model and incorporate richer semantic context by means of topological graph structures (Stoehr et al., 2019, 2020).

References

Richard W. Brislin. 1970. Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1(3):185–216.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

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.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

P.J. Finlay. 2021. Argos translate. Open source neural machine translation software.

Lukas Gebhard and Felix Hamborg. 2020. The polusa dataset: 0.9 m political news articles balanced by time and outlet popularity. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 467–468.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ali Hürriyetoğlu, Osman Mutlu, Farhana Ferdousi Liza, Erdem Yörük, Ritesh Kumar, and Shyam Ratan. 2021. Multilingual protest news detection - shared task 1, case 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Çağrı Yoltar, Burak Gürel, Fırat Duruşan, and Osman Mutlu. 2019a. A task set proposal for automatic protest information collection across multiple countries. In *Advances in Information Retrieval*, pages 316–323, Cham. Springer International Publishing.

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Çağrı Yoltar, Burak Gürel, Fırat Duruşan, Osman Mutlu, and Arda Akdemir. 2019b. Overview of clef 2019 lab protestnews: Extracting protests from news in a cross-context setting. In *Experimental IR*

Meets Multilinguality, Multimodality, and Interaction, pages 425–432, Cham. Springer International Publishing.

Ali Hürriyetoğlu, 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).

Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Fırat Duruşan, Çağrı Yoltar, Deniz Yüret, and Burak Gürel. 2021. Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction. *Data Intelligence*, pages 1–28.

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 preprint arXiv:1907.11692*.

Nikola Mrkšić, Diarmuid O Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. *arXiv preprint arXiv:1603.00892*.

Arpit Rajauria. 2020. Pegasus fine-tuned for paraphrasing. Open source neural machine translation software.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Niklas Stoehr, Fabian Braesemann, Michael Frommelt, and Shi Zhou. 2020. Mining the automotive industry: A network analysis of corporate positioning and technological trends. In *Complex Networks XI*, pages 297–308. Springer International Publishing.

Niklas Stoehr, Marc Brockschmidt, Emine Yilmaz, and Jan Stühmer. 2019. Disentangling interpretable generative parameters of random and real-world graphs. In *arXiv*, volume 1910.05639.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *arXiv preprint arXiv:1509.01626*.

A Appendix

A.1 Second Pretraining Considerations

Gururangan et al. (2020) propose two types of additional pretraining: domain-adaptive pretraining (DAPT) and task-adaptive pretraining (TAPT). DAPT involves a *second pretraining* on large corpus of text from a specific domain (e.g news paper articles) while TAPT uses unlabeled training data for the downstream task. We consider the second-pretrained DAPT and TAPT model for AG News (Zhang et al., 2015) and finetune them for our task of event sentence classification. The results are shown in Table 6. We see that the F1 score of the DAPT model is almost 0.01 lower than the finetuned RoBERTa default model and TAPT performs even worse. We believe that training on a general news corpus would not help improve the embedding quality for our task because the AG News dataset contains articles of different categories (e.g business, technology and sports) while our subtask only deals with political news. This is consistent with our observation in Section 5.2 when we see worse result as we add negative data from Subtask 1. Ideally, we would perform TAPT using unlabeled training data, which involves protest news articles from Indian Express, New Indian Express, The Hindu, Times of India, South China Morning Post, and People's Daily, according to (Hürriyetoğlu et al., 2021). This would ensure no domain gap between our data for *second pretraining* and finetuning. Due to time and resource constraint, however, we cannot gain access to articles from these outlets. Thus, we resort to POLUSA (Gebhard and Hamborg, 2020). While it is not from the same outlets, the fact that it only contains political news make it suitable for our purpose.

Initialization	F1 Score
RoBERTa	**0.8283**
DAPT	0.8195
TAPT	0.8155

Table 6: Validation performance of finetuning DAPT and TAPT models, second pretrained on AG News, compared to the finetuned RoBERTa default model

A.2 Data Augmentation

In this section, we discuss the four different data augmentation methods we consider in the main paper.

Back-translation Back-translation means translating the source text into a different language, and translate back to the source language. This method have been used since the 1970s in translation quality research (Brislin, 1970) and have recently been used to improve machine translation models (Sennrich et al., 2015; Edunov et al., 2018). In our implementation, we use Chinese as the intermediate language.

Embedding Augmentation Embedding Augmentation performs augmentation by replacing words with neighbors in the counter-fitted embedding space (Mrkšić et al., 2016).

Checklist Augmentation Checklist Augmentation is based on Ribeiro et al. (2020). This method augments texts by replacing names, locations, and numbers detected in the text as well as performing contraction and extension.

Paraphrasing Paraphrasing refers to augmenting text by generating a paraphrased version of the text. We use the Pegasus (Zhang et al., 2020) model finetuned for paraphrasing (Rajauria, 2020) to generate paraphrased text. We include both the original training set and the paraphrased training set for finetuning our model.

Example We show example sentences of data augmentation methods in Table 7. We see that back-translation, checklist and embedding augmentation perform their intended functions, while paraphrasing seems to create facts that are not present in the original sentence.

A.3 Other Finetuning Setting

We explore other finetuning settings and show the results in Table 8. This experiment is done using the RoBERTa default weight without *second pretraining*. We consider the following setting: 1. We add two more fully connection (FC) layers before the output. We still finetune the entire model; 2. We consider setting 1 but freeze the RoBERTa backbone; 3. We consider the output of all tokens in the last hidden layer, and pass them through an LSTM (Hochreiter and Schmidhuber, 1997) layer before the classification head; 4. Setting 3 but with frozen RoBERTa backbone. We see that more FC layers does not help, and that when we only consider the [CLS] embedding, freezing the main model would result in very bad performance. At the same time, when we consider embeddings from all tokens with

Method	Sentence
Original Sentence	"Purandeswari, who on Tuesday said when it was certain that Telangana would be a reality there was no point in demanding something that was not going to be delivered, reiterated her new stance on Wednesday."
Back-translation	"Tuesday said that Prandeswari was aware that Teangana would be a reality without any requirement to do so, and she therefore reiterated her new position on Wednesday."
Checklist + Embedding	"Mareli, who on Tuesday said when it was certain that Telangana would be a reality there was no point in demanding something that was not going to be delivered, reiterated her new stance on Wednesday."
Paraphrasing	"Thousands of students are writing their National Senior Certificate (matric) exams and could fail to arrive on time."

Table 7: Example sentences from each data augmentation method that we consider: back-translation, embedding augmentation, checklist augmentation and paraphrasing.

an LSTM layer, we get a small boost in performance. Freezing the main model does not hurt nearly as much in this setting, suggesting a possible way of finetuning large language models in resource-constrained situations. Given that using embeddings from all tokens is not the conventional setup of a RoBERTa model for downstream classification tasks, we still use the conventional setting by connecting the [CLS] embedding to an FC layer in our submission.

Setting	F1 Score
Default	0.8283
1	0.8280
2	0.5631
3	**0.8301**
4	0.8155

Table 8: Performance of the model under other finetuning settings. Setting Default: the standard way - RoBERTa model with a FC layer connected to [CLS] embedding for classsification. Setting 1: two more fully connection (FC) layers before the classification head. Setting 2: Setting 1 with the backbone model frozen. Setting 3: Pass embeddings of all tokens to an LSTM before the output layer. Setting 4: Setting 3 with the backbone model frozen.