NSUT-NLP at CASE 2022 Task 1: Multilingual Protest Event Detection using Transformer-based Models

Manan Suri, Krish Chopra, Adwita Arora
Netaji Subhas University of Technology, New Delhi
{manan.suri.ug20, krish.ug20, adwita.ug20} @nsut.ac.in

Abstract

Event detection, specifically in the socio-political domain, has posed a long-standing challenge to researchers in the NLP domain. Therefore, the creation of automated techniques that perform classification of the large amounts of accessible data on the Internet becomes imperative. This paper is a summary of the efforts we made in participating in Task 1 of CASE 2022. We use state-of-art multilingual BERT (mBERT) with further fine-tuning to perform document classification in English, Portuguese, Spanish, Urdu, Hindi, Turkish and Mandarin. In the document classification subtask, we were able to achieve F1 scores of 0.8062, 0.6445, 0.7302, 0.5671, 0.6555, 0.7545 and 0.6702 in English, Spanish, Portuguese, Hindi, Urdu, Mandarin and Turkish respectively achieving a rank of 5 in English and 7 on the remaining language tasks.

1 Introduction

Protests exist as a natural way for citizens of a nation to show their dissatisfaction with decisions taken by the respective governments or authorities (Neogi et al., 2021). The sentiment prevalent in such events and the reaction by various parties to these events provide the basis for carrying out many studies in the sociopolitical field, such as the public opinion about the event that was the cause for the protest, how much freedom the protesters were afforded as a measure of the democracy in the nation, and so on. With the advancement of technology, there has also been an exponential rise in the use of social networks as a medium for exchanging information across the globe, with global events and their inner nuances being available to the public at large. However, extracting valuable insights from such events on a national or global scale is a daunting task if done manually (Carothers and Youngs, 2015). Even if we leverage automated techniques for the process, there are numerous challenges faced while working on multilingual data (Hershcovich et al., 2022). Hence, there exists an incentive to automate the task of processing protest news from multiple locations and in multiple languages and to create an NLP system that could be generalized for the task of detecting protest news. Task 1 of CASE 2022 (Hürriyetoğlu et al., 2022a) aims at working on multilingual protest news corpora, with Subtask 1 working towards the binary classification of news reports, where if a document reports on an event that has happened or is ongoing, it is marked as relevant, otherwise it is considered irrelevant.

Our approach revolves around the use of state-of-art Pre-Trained Language Models (PLMs) and finetuning them to perform the task we require. We leverage the `bert-base-multilingual-cased` (Devlin et al., 2018) that was trained in over 104 languages to tackle the multilingual task. We finetune it for protest news detection. Since most of the training datasets had a bias toward the negative class, we augmented the datasets by translating positive samples from other language datasets and hence improving the balance between the positive and negative class to prevent our model from being biased towards the negative class. Furthermore, the lack of samples in Portuguese and Spanish presents us with a few-shot learning scenario, which we tackle by augmenting these datasets with samples from the English dataset translated into the respective languages. For languages with no training datasets (Urdu, Turkish, Mandarin and Hindi), we created training datasets by translating the English corpus.

The rest of the paper is organised as follows: We begin by laying out the past literature and work done in the field of protest event detection in Section 2 followed by the description of the task at hand and the data given to us in Section 3. In Section 4, we describe the techniques we employed, namely data augmentation and the model we used, multilingual BERT (mBERT). The experimental

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 161 - 168
December 7-8, 2022 ©2022 Association for Computational Linguistics

setup for our system is described in Section 5 and the results on the test set are mentioned and analysed in Section 6. Finally, in Section 7, we draw a conclusion to our work and go over prospective directions for additional research.

2 Related Work

Protest detection and allied fields have drawn a lot of attention from researchers in the NLP domain. MAVEN (Wang et al., 2020) and CySecED (Trong et al., 2020) are annotated datasets in the English language created for the purposes of event detection. ACE 2005 (Walker, Christopher et al., 2006) and TempEval-2 (Verhagen et al., 2010) are multilingual datasets where ACE 2005 covers English, Arabic, and Chinese and TempEval-2 covers Chinese, English, French, Italian, Korean and Spanish. MINION (Veyseh et al., 2022) is another multilingual ED dataset covering 8 different languages (English, Spanish, Portuguese, Polish, Turkish, Hindi, Japanese and Korean). MM-CHIVED (Steinert-Threlkeld and Joo, 2022) is another dataset containing multimodal data like text and images compiled from social media regarding Chile and Venezuela protests. There have also been region-specific case studies, such as detection of protest events in Turkey 2013 (Elsafoury, 2020) and protest analysis in Greece over the last twenty years through the scope of Computational Social Science (Papanikolaou and Papageorgiou, 2020). Previously event detection has also been researched upon by researchers participating in the Task 1 of CASE 2021 (Hürriyetoğlu et al., 2021). Teams which participated in the task earlier have used multilingual pre-trained language models (Re et al., 2021; Awasthy et al., 2021a; Gürel and Emin, 2021) which is similar to the approach used by our system.

3 Background

3.1 Task

Event Detection aims at extracting event triggers (in the forms of singular nouns or verbs or even full sentences sometimes) and classifying the triggers into the type of event they belong to (Awasthy et al., 2021b). The main challenge of this task comes from the fact there exists a many-to-many relationship between the trigger and event type, i.e. the same event can be represented by various event triggers and the same expression can represent different events in different contexts (Feng et al., 2016). The CASE 2022 workshop (Hürriyetoğlu et al., 2022a) focuses on protest news event detection. In this paper, we aim to tackle Shared Task 1: Multilingual Protest News Detection, specifically Subtask 1.

Subtask 1 - Document Classification is a binary classification task on the document-level (news article) where we classify an event as positive if the event actually occurred or is ongoing. Scheduled events, rumors, and speculations are considered as irrelevant and hence marked as negative.

The task we deal with is a binary classification task where we classify documents that pertain to ongoing or already occurred events as positive samples. Events that are merely rumors, scheduled to take place in future or speculations are marked as negative samples.

The task is multilingual, as we have training data consisting of English, Portuguese, and Spanish Languages for both training and evaluation of the model. The Portuguese and Spanish datasets present us with a few-shot scenario to the dearth of data compared to the English data set. At the same time, Hindi, Mandarin, Turkish and Urdu evaluation sets present a zero-shot setting to evaluate our model.

The metric used for the evaluation of the results produced by the model is the Macro-F1 score. It provides a balance between Precision and Recall of the model, by taking a harmonic mean of both metrics.

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

3.2 Data

Language	Split	Subtask 1
English	Train	9,324
	Test	3,871
Spanish	Train	1,000
	Test	671
Portuguese	Train	1,487
	Test	671
Hindi	Test only	268
Urdu	Test only	299
Turkish	Test only	300
Mandarin	Test only	300

Table 1: Distribution of samples in respective datasets for the given languages

	English	Spanish	Portuguese
Characters	1067.37	932.87	635.25
Tokens	199.74	177.35	116.03

Table 2: Analysis of the average number of characters and tokens in the respective training sets for English, Spanish, and Portuguese.

The data for this task has been created and annotated using the methods described in Hürriyetoğlu et al. (2022b). While the task is multilingual, there isn't an even distribution of data for all languages, with the English corpus having more data than both Portuguese and Spanish. Also, some languages to be evaluated do not have any training data (a zero-shot learning problem), namely Hindi, Turkish, Urdu and Mandarin. The distribution of data is given in Table 1 as shown.

The distribution of labels for training data for subtask 1 is as follows: The positive sample ratio for Subtask-1 is 0.205 for English dataset, 0.131 for Spanish dataset and 0.132 for Portuguese.

This highlights that the data is skewed towards the negative class for all languages. It is natural to tackle this bias problem so that our model does not align itself too much with one class, which would lead to its performance suffering in a more balanced scenario.

The number of characters in each training dataset is shown in Table 2. One would believe that a longer sentence gives the model more context to work with and therefore produces better results; however, a longer text also runs the risk of confusing the model with interference from mixed signals (Çelik et al., 2021). The number of tokens in each language dataset is also shown.

Another thing to note is the low amount of training data in the case of Portuguese and Spanish, and the complete lack of it in the case of Hindi, Urdu, Mandarin and Turkish. We attempt to alleviate this problem by translating the English corpus examples into the respective language and training on this augmented dataset.

4 System Overview

4.1 Data Augmentation

Data augmentation refers to the set of techniques to increase the quantity and diversity of data points in a data-set without collecting new data. The purpose of data augmentation in our system was as follows:

- **Class Imbalance:** In the English, Spanish and Portuguese datasets provided by the organizers, the ratio of the positive samples was 0.205, 0.131 and 0.132 for Subtask 1. Therefore to provide enough diversity of samples of the positive class, data augmentation was required

- **Lack of Training Data:** Spanish and Portuguese had limited training data compared to English. For Hindi, Urdu, Mandarin and Turkish, no training data was available. Therefore to create an appropriately large dataset, data augmentation is used.

The technique used for data augmentation in our system leverages the availability of three linguistically different datasets. We translated various combinations of positive and negative samples from the three available datasets of English, Spanish and Portuguese

Our augmentation strategy can be understood by Fig 1. The process is described below:

1. **English** The final training set consisted of the original English dataset along with positive samples of Spanish and Portuguese datasets translated into English.

2. **Spanish** The final training set consisted of the original Spanish dataset along with the English dataset (both positive and negative samples), Portuguese dataset translated into Spanish.

3. **Portuguese** The final training set consisted of the original Portuguese dataset along with the English dataset (both positive and negative samples), Spanish dataset translated into Spanish

4. **Hindi, Urdu, Mandarin and Turkish** The training datasets for Hindi, Urdu, Mandarin and Turkish were created by translating the final English dataset into the respective languages.

Table 3 displays the size and final data distribution of the respective train datasets after data augmentation.

4.2 Finetuning Pretrained MultiModal BERT

Pre-training in NLP refers to moulding a large collection of unannotated text input into general-purpose language representations. It is useful as it

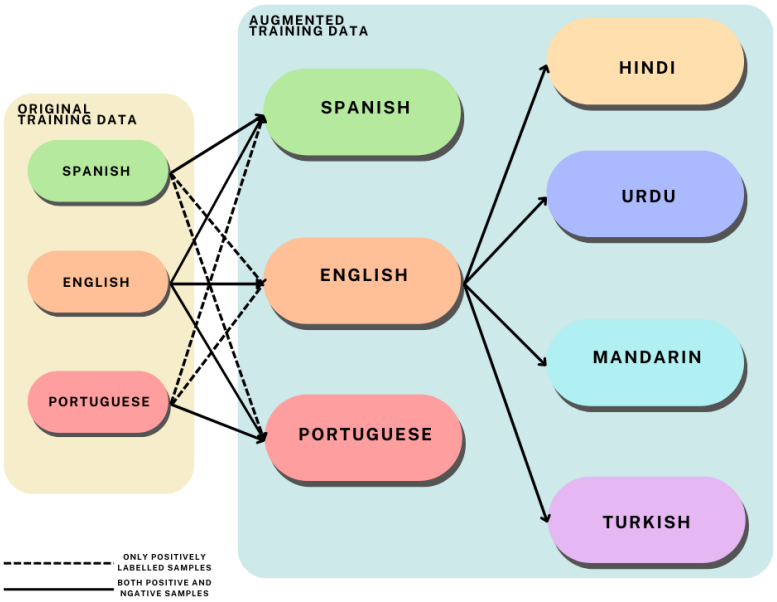

Figure 1: Data augmentation using translation. In the given diagram, each directed edge represents translation from a source language to a target language. The graph represents the combination of datasets used during translation to achieve the final augmented training sets.

Language(s)	Label 0	Label 1	Total
English, Hindi, Urdu, Mandarin and Turkish	2240	7412	9652
Spanish	2240	8281	10521
Portuguese	2240	8702	10942

Table 3: Distribution of labels in the respective training set after data augmentation.

prevents having to start from scratch when training a new model for downstream tasks. Because it offers a stronger model initialization, pre-training improves generalization performance and aids in convergence on downstream tasks. Pretraining can be considered a form of regularization that avoids overfitting on smaller datasets with relatively few human-annotated examples. On many NLP tasks, pre-training models followed by fine-tuning them for downstream tasks, have demonstrated good performance (Erhan et al., 2010).

The model used in our system is based on the BERT architecture. Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is a transformer-based (Vaswani et al., 2017) pre-trained language model that was created with the objective of fine-tuning a pre-trained model yields better performance. The pretraining phase of BERT includes two tasks. Firstly, Masked Language Modeling (MLM) is where certain words are randomly masked in a sequence. About 15% of the words in a sequence are masked. The model then attempts to predict the masked words. Secondly, Next Sentence Prediction (NSP), where the model has an additional loss function, NSP loss, indicates if the second sequence follows the first one. Around 50% of the inputs are a pair, and they randomly chose the other 50.

Our system uses a multimodal BERT (mBERT) specifically, `bert-base-multilingual-cased` which has been trained on 104 languages with the largest Wikipedia content. Since the size of Wikipedias for different languages varies, exponentially smoothed weighting of the data is performed to under-sample resource-rich languages and over-sample low-resource languages. The model has 12 layers of transformer blocks with 768 hidden dimensions conditioned on 12 self-attention heads. In total, the model has 110M trained parameters.

Preprocessing involves splitting the input document into tokens and generating a compatible in-

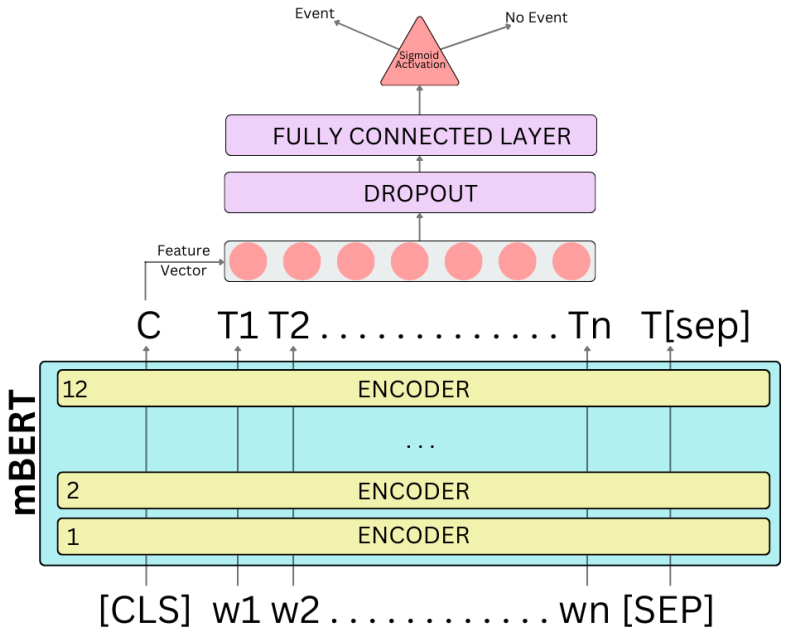

Figure 2: Diagrammatic representation of the model used for the system.

put sequence. Considering that different languages have different vocabularies, the model uses a shared 110k WordPiece vocabulary. For Mandarin texts, a whitespace is inserted around every character before applying the WordPiece tokenizer, making the Mandarin input character tokenized. For other languages, lower casing and accent removal is the first step. This is followed by punctuation splitting and finally whitespace tokenization. Special tokens, [CLS] used to indicate the beginning of the input, [SEP] used to indicate the end of a sequence and [PAD] used for padding sequences to max-length are inserted into the tokenized sequence. Fine-tuning of the model involved stacking a dense layer on top of the BERT output. The dense layer is stacked with a dropout layer. The final layer of the model consists of two neurons with sigmoid activation to predict the binary labels.The features of the [CLS] token are used for classification. A benchmark of 0.62 was used to classify a sample as positive. Fig 2 summarises the model architecture used by our system.

5 Experimental Set-up

The models were developed on Keras[1] (Chollet et al., 2015), and implemented using the transformers library by HuggingFace[2] (Wolf et al., 2019). The model used is `bert-base-multilingual-cased` [3]. We use the AutoTokenizer [4] offered by HuggingFace's transformers library to tokenize our inputs. We experimented with learning rates of 1e-5, 3e-5 and 5e-5 for all models, finding the best results at 3e-5. For all the models, we fixed the max length parameter at 512 tokens and the batch size parameter to 6. The finetuning for the models was performed on Google Colab GPU. We trained each model for 3-4 epochs and found the best results at 4 epochs. The dropout rate during fine-tuning is 0.2. We used the Adam (Kingma and Ba, 2014) optimizer from Keras. The loss function used is binary cross-entropy. The translation was performed using the Google Translation library in Python `googletrans(v3.1.0a0)` [5].

6 Results and Discussion

Table 4 demonstrates the results of our system on the test set for the respective languages. One common pitfall of the system across languages is that it performs better on the majority class and fails to identify the minority class correctly. Our hypothesis is that this happens because despite data augmentation increasing the count of samples, the dataset is still imbalanced. The quality of aug-

[1] https://keras.io/

[2] https://huggingface.co/docs/transformers/index

[3] https://huggingface.co/bert-base-multilingual-cased

[4] https://huggingface.co/transformers/v3.0.2/model_doc/auto.html#autotokenizer

[5] https://pypi.org/project/googletrans/

mented samples depends on the performance of the translation engine which is a decisive factor in our system. Furthermore, we believe that a task like protest event detection involves nuanced references and linguistic nuances may get lost during translation, even more so when the datasets for Hindi, Urdu, Mandarin and Turkish are generated through two cycles of translation.

Our model seems to have performed better on the English dataset indicating that the multilingual BERT has a better contextualizing ability for the *Lingua Franca*, English. The preprocessing process involves removal of accents which might be detrimental to performance of many languages which heavily rely on accents such as Hindi, Urdu, Turkish and Spanish. For example, in Turkish ı and i (non -dotted and dotted) are very different vowels with the phonetic sounds (as in cycle - sīkl) and **ē** (as in easy - **ēzē**).

Language	Macro F1 Score
English	0.8062
Spanish	0.6445
Portuguese	0.7302
Hindi	0.5671
Urdu	0.6555
Mandarin	0.7545
Turkish	0.6702

Table 4: The results on the given test set for each of the respective languages given by our system. The metric for evaluation is the Macro F1 score.

7 Conclusion and Future Work

The amounts of publicly available data on the Internet, especially social networks, desire for skillful analysis for the purposes of protest detection. This becomes especially imperative because of the significance of protests in the social, political and economic domains. Our submission in Task 1 of CASE 2022 demonstrated the effective use of Pretrained Language Models (PLMs), specifically multilingual BERT (mBERT) in the binary classification of documents into events or not events. We were also successful in tackling the dearth in training data and class imbalance using data augmentation. We have been able to achieve F1 scores of 0.8062, 0.6445, 0.7302, 0.5671, 0.6555, 0.7545 and 0.6702 in English, Spanish, Portuguese, Hindi, Urdu, Mandarin, and Turkish respectively. In the future, we can deal with class imbalance using class weighing (Suri, 2022). We would also like to experiment with cross-lingual finetuning on a multilingual model by training in one language and testing in another language. We would like to extend this work by using language specific PLMs rather than a multilingual model.

References

Parul Awasthy, Jian Ni, Ken Barker, and Radu Florian. 2021a. IBM MNLP IE at CASE 2021 task 1: Multigranular and multilingual event detection on protest news. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 138–146, Online. Association for Computational Linguistics.

Parul Awasthy, Jian Ni, Ken Barker, and Radu Florian. 2021b. IBM MNLP IE at CASE 2021 task 1: Multigranular and multilingual event detection on protest news. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 138–146, Online. Association for Computational Linguistics.

Thomas Carothers and Richard. Youngs. 2015. The complexities of global protests.

Furkan Çelik, Tuğberk Dalkılıç, Fatih Beyhan, and Reyyan Yeniterzi. 2021. SU-NLP at CASE 2021 task 1: Protest news detection for English. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 131–137, Online. Association for Computational Linguistics.

Francois Chollet et al. 2015. Keras.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

Fatma Elsafoury. 2020. Teargas, water cannons and twitter: A case study on detecting protest repression events in turkey 2013. In *Text2story@ ecir*.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, pages 625–660.

Xiaocheng Feng, Lifu Huang, Duyu Tang, Heng Ji, Bing Qin, and Ting Liu. 2016. A language-independent neural network for event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 66–71, Berlin, Germany. Association for Computational Linguistics.

Alaeddin Gürel and Emre Emin. 2021. ALEM at CASE 2021 task 1: Multilingual text classification on news articles. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 147–151, Online. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural nlp.

Ali Hürriyetoğlu, Osman Mutlu, Fırat Duruşan, Onur Uca, Alaeddin Selçuk Gürel, Benjamin Radford, Yaoyao Dai, Hansi Hettiarachchi, Niklas Stoehr, Tadashi Nomoto, Milena Slavcheva, Francielle Vargas, Aaqib Javid, Fatih Beyhan, and Erdem Yörük. 2022a. Extended multilingual protest news detection - shared task 1, case 2021 and 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 (ACL).

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, Hristo Tanev, Vanni Zavarella, Reyyan Yeniterzi, Osman Mutlu, and Erdem Yörük. 2022b. Challenges and applications of automated extraction of socio-political events from text (case 2022): 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 2022)*, online. Association for Computational Linguistics (ACL).

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Ashwin Sanjay Neogi, Kirti Anilkumar Garg, Ram Krishn Mishra, and Yogesh K Dwivedi. 2021. Sentiment analysis and classification of indian farmers' protest using twitter data. *International Journal of Information Management Data Insights*, 1(2):100019.

Konstantina Papanikolaou and Harris Papageorgiou. 2020. Protest event analysis: A longitudinal analysis for greece. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 57–62.

Francesco Re, Daniel Vegh, Dennis Atzenhofer, and Niklas Stoehr. 2021. Team "DaDeFrNi" at CASE 2021 task 1: Document and sentence classification for protest event detection. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 171–178, Online. Association for Computational Linguistics.

Zachary Steinert-Threlkeld and Jungseock Joo. 2022. Mmchived: Multimodal chile and venezuela protest event data. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1332–1341.

Manan Suri. 2022. PiCkLe at SemEval-2022 task 4: Boosting pre-trained language models with task specific metadata and cost sensitive learning. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 464–472, Seattle, United States. Association for Computational Linguistics.

Hieu Man Duc Trong, Duc-Trong Le, Amir Pouran Ben Veyseh, Thut Nguyn, and Thien Huu Nguyen. 2020. Introducing a new dataset for event detection in cybersecurity texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5381–5390.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Nguyen. 2022. Minion: a large-scale and diverse dataset for multilingual event detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2286–2299.

Walker, Christopher, Strassel, Stephanie, Medero, Julie, and Maeda, Kazuaki. 2006. Ace 2005 multilingual training corpus.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. Maven: A massive general domain event detection dataset. *arXiv preprint arXiv:2004.13590*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.