ALEM at CASE 2021 Task 1: Multilingual Text Classification on News Articles

Alaeddin Selçuk Gürel
selcuk.gurel@bilgiedu.net

Emre Emin
emreemin@sabanciuniv.edu

Abstract

We participated CASE shared task in ACL-IJCNLP 2021. This paper is a summary of our experiments and ideas about this shared task. For each subtask we shared our approach, successful and failed methods and our thoughts about them. We submit our results once for every subtask, except for subtask3, in task submission system and present scores based on our validation set formed from given training samples in this paper. Techniques and models we mentioned includes BERT, Multilingual BERT, oversampling, undersampling, data augmentation and their implications with each other. Most of the experiments we came up with were not completed, as time did not permit, but we share them here as we plan to do them as suggested in the future work part of document.

1 Introduction

This paper includes review and explanations about our ideas and experiments for the CASE shared task in ACL-IJCNLP 2021. The main purpose and goal for this shared task is to identify and classify sociopolitical and crisis event information at multiple levels and languages.

Main categories for subtasks are document classification (subtask1), sentence classification (subtask2), event sentence coreference identification (subtask3) and event extraction (subtask4). Each subtask has three batches of training data which are in English, Spanish and Portuguese (Hürriyetoğlu et al., 2020, 2019a,b).

Document classification and sentence classification tasks are binary classification tasks which aim to classify news articles and sentences respectively. The classification criteria of the document classification task is whether news article contains at least one past or ongoing event. Sentence classification is also a binary classification task, sentences are labeled as 1 if they contain event triggers within them. Event sentence coreference identification task aims to identify which event sentences are referring the same event. The objective of the event extraction task is to gather event trigger information and event information from given news article.

We participated in subtask1, subtask2 and subtask4. The training data for subtask3 was not sufficient for us to build and optimize the model for the given time schedule, since it was not possible to get exact results for test data. Our results are based on validation data that we constructed from given training data.

We propose a multilingual BERT Model (Devlin et al., 2018) for the shared task 1 (Hürriyetoğlu et al., 2021a,b). We trained and measured the performance of our model which is fine-tuned in English, Spanish and Portuguese. The model is formed by using and modifying multiple pretrained BERT models for each subtask and language we participated for [1].

2 Data

2.1 Training Data

Training data includes three languages for each subtask, English, Spanish and Portuguese. The data distributions are given below for each level. For both document classification and sentence classification tasks, training data was shared in JSON Lines text format. In this data, each document/sentence has an ID, text and label. The data of event extraction task was shared in similar format to CoNLL format. In token level data, documents are starting with SAMPLE_START token, document and sentences are separated by empty lines and [SEP] token respectively.

[1] Code that we used for this shared task submission can be found at https://github.com/alaeddingurel/ALEM-CASE2021

Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) , pages 147–151
August 5–6, 2021, ©2021 Association for Computational Linguistics

Language	0	1	Total
English	7412	1912	9324
Spanish	802	198	1000
Portuguese	1184	303	1487

Table 1: Label distribution of training data in document level

The total number of documents, sentences and tokens provided for the English Language was much larger than other source languages.

Language	0	1	Total
English	18602	4223	22825
Spanish	2232	509	2741
Portuguese	961	221	1182

Table 2: Label distribution of training data in sentence level

There are seven different categories in event extraction dataset which are etime (Event time), fname (Facility name), organizer, participant, place, target and trigger.

3 Methodology

We used Huggingface's transformers (Wolf et al., 2020) library in order to fine-tune our BERT model for each subtask. We fine-tuned separate BERT models, each model pre-trained using a corpus in their respective language. The training data provided was quite unbalanced for every language in terms of both sample size and label distribution. We have tried over and under sampling techniques using imbalanced-learn package (Lemaître et al., 2017) to form a better training split. Both of the methods for our case affected the results in a negligible amount. So we decided to use naive random sampling for our experiments.

One other obstacle we worked on is BERT's maximum token size for its inputs. Tokenized input given to BERT is trimmed if it includes more than 512 tokens. This is a huge data loss for our subtasks, especially for document level classification. Many documents are trimmed by default configuration, so we tried a populating method to avoid losing any data with cost of extra labelling process. The idea is to split the data to be trimmed into chunks less than 512 tokens and label each one as it was labeled before splitting. This may cause a incorrect labeling process since the document is now cut into texts and each one of them may be against its parent label by its own in the training process. As a practical example of this method, let's say we have a text $Z = X_0 \cdot X_1 \cdot ... \cdot X_n$, where each X_i are strings that form Z when concatenated. Tokenized length of Z is greater than 512 and it is labeled as 0 in training set. We split Z into X_is to obtain less than 512 tokens for each part and set the labels of each X_i as 0. This blind labelling process may cause incorrectly assigned labels for some X_is, since label 1 may be more suitable for their individual meanings. However we did not observe a significant change on the results for any of the languages. Considering this method did not improve the results, we did not use it for our final tests.

We also used this method in the prediction phase. The texts were splitted similarly as in the given example. The final prediction was decided by majority of votes method e.g. if 3 texts are labeled as 1,1,0, then their parent prediction is 1 as it has higher vote.

- English - BERT
- Spanish - BETO (Cañete et al., 2020)
- Portuguese - BERTimbau Base (Souza et al., 2020)

For the multilingual BERT experiments we have used the pretrained mBERT model in order to fine-tune our data for subtasks. We used BERT tokenizer which is based on WordPiece tokenization algorithm. We splitted training data with the purpose of forming a test set before submitting the final results to shared task system. The split for train and test data distributed 80% to 20% respectively. The method we use concatenates all English, Spanish and Portuguese data and train them altogether. The split is deterministic and stayed same for all of our experiments for all models in order to obtain results for the same test data.

4 Experiments

The scores we demonstrate on the document classification and sentence classification are based on f1-macro metric. The evaluation criteria that we used in event extraction for validation data is f1 score. We experimented with various epoch numbers and batch sizes with the intent of optimizing the hyper-parameters. We made our decisions to use these epoch numbers and batch sizes based on

Language	etime	fname	organizer	participant	place	target	trigger
English	1209	1201	1261	2663	1570	1470	4595
Spanish	40	49	25	88	15	64	157
Portuguese	41	48	19	73	61	32	122

Table 3: Label distribution of training data in token level

our experimental setup. The epoch and batch parameters given to training phase for BERT Base for document classification task with epoch as 5 and batch as 32, sentence classification task with epoch as 3 and batch as 64. For Multilingual BERT we fine-tuned parameters as 3 epochs and 32 batches for document classification task and 5 epochs and 32 batches for sentence classification task.

Language	mBERT	BERT
English	84.17%	**84.26%**
Spanish	**76.32%**	73.82%
Portuguese	79.78%	**80.20%**

Table 4: Results for document level

English BERT gives better results in comparison with multilingual BERT model by 0.09%. In our experiments we observed that multilingual BERT model has superior results for Spanish Language by 2.5% when compared to Spanish BERT model used in terms of our measurement criteria. Portuguese BERT has a higher f1-macro score by 0.42% when we compare it with its counterpart, multilingual BERT. There is no significant gap between the f1-macro scores of multilingual BERT and BERT Base models which are pretrained with their respective languages.

Language	mBERT	BERT
English	84.70%	**87.68%**
Spanish	76.53%	**83.95%**
Portuguese	82.01%	**82.72%**

Table 5: Results for sentence level

BERT models pretrained with respective languages has greatest scores with comparison with multilingual BERT for all languages in sentence classification task.

Token	**f1-Score**
etime	77.95
fname	54.65
organizer	65.89
participant	75.40
place	83.86
target	55.10
trigger	84.32

Table 6: Results for token level

There isn't enough data points for Spanish and Portuguese languages for training and evaluation of event extraction task. We think that we need different approaches in order to train and evaluate this data for further testing, but we share the evaluation performance results for English language since it has enough data points to form an acceptable model when compared to the other languages. We made our document, sentence and event extraction submissions based on BERT base models which are trained with their respective languages for each . We used f1-score metric with the purpose of analysing event extraction performance for each token category.

5 Conclusion and Future Work

This paper describes our system description for submission for CASE @ ACL-IJCNLP 2021: Socio-Political and Crisis Events Detection shared task. In training phase, we performed our experiments using separate pretrained language models with different training data. We report their performance for 3 tasks with the addition of the results for multilingual BERT model. We also compared our models with the other BERT models which are trained with their respective language data. We tested our fine-tuned language models with the test data provided by shared task organizers and made our submissions for document classification and sentence classification tasks. We achieved 80.82, 72.98 and 46.47 f1-macro scores in document classification. f1-macro scores of the sentence classification task are 79.67, 42.79 and 45.30 for English, Portuguese and Spanish respectively. We didn't make submission for token classification task due time limitations, but shared the results we observed in tests on our validation set.

One of the important issues with BERT is to optimize the training data in order to align with its maximum token size while training. In some tasks, especially in document level classification, this is a significant factor for pre-processing, since the length of the input texts are too long for being tokenized to fit BERT as whole. This situation leads to an experiment devoted for managing this limitation.

Following our experiments in over- and under-sampling methods, we would like to use data augmentation for future training methods in order to achieve an equilibrium in terms of training data labels. Augmenting method may be text generation from already given documents and sentences, but we do not expect this method being successful for languages other than English since our sample data is not as much for the other languages.

One another method we considered applying for future experiments was ensemble learning. The idea is training different models for the same task and observe their differentiated scores and group them by their success on predicting particular inputs. This method has a cost of training many models and measuring their prediction success with respect to the others, however after forming an optimal set of models, we can use them to unite on a cumulative score on a single input by assigning a weight for each of their individual output. This idea of combining many models can be also used for BERT initiated environment by constructing a system where the structure is built on top of BERT and inserting custom networks into its embedding layers.

There are many improvements and analysis to be done in order to understand strengths and weaknesses of this system and further improvements might be added on top of it.

References

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Ali Hürriyetoğlu, Osman Mutlu, Farhana Ferdousi Liza, Erdem Yörük, Ritesh Kumar, and Shyam Ratan. 2021a. 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, Jakub Piskorski, Reyyan Yeniterzi, and Erdem Yörük. 2021b. 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)*, online. Association for Computational Linguistics (ACL).

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Osman Mutlu, Çağrı Yoltar, Fırat Duruşan, and Burak Gürel. 2020. Cross-context news corpus for protest events related knowledge base construction. In *Automated Knowledge Base Construction*.

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.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.