CEIA-NLP at CASE 2022 Task 1: Protest News Detection for Portuguese

Diogo Fernandes Costa Silva, Adalberto Junior, Gabriel da Mata Marques, Anderson da Silva Soares and Arlindo Rodrigues Galvao Filho
Federal University of Goiás
Goiânia, Goiás
diogo_fernandes@discente.ufg.br

Abstract

This paper summarizes our work on the document classification subtask of Multilingual protest news detection of the CASE @ ACL-IJCNLP 2022 workshop. In this context, we investigate the performance of monolingual and multilingual transformer-based models in low data resources, taking Portuguese as an example and evaluating language models on document classification. Our approach became the winning solution in Portuguese document classification achieving 0.8007 F1 Score on Test set. The experimental results demonstrate that multilingual models achieves best results than monolingual models in scenarios with few dataset samples of specific language, because we can train models using datasets from other languages of the same task and domain.

1 Introduction

Observing the prominent ease of use and variety of virtual media, such as social networks in general, and the exponential use of these for the organizational purpose of various manifestations, protests and social movements (McKeon and Gitomer, 2019), a large amount of information is stored in databases of applications that are not properly analyzed for a socially beneficial purpose. Therefore, it is important to explore alternatives for an analysis capable of classifying and even predicting the organization of social movements such as those mentioned above.

Considering the importance of detecting crises and sociopolitical events present in social networks (Hürriyetoğlu et al., 2022). The practical application of extracting and classifying information and its importance in the field of collective social manifestations, in order to obtain several useful results for important political and economic decisions (Duruşan et al., 2022).

In this paper, we investigate the performance of monolingual and multilingual language models for classification of documents in Portuguese. The experiments are conducted on Socio-political datasets and all models are transformer-based models. Our submission achieved the 1st place in document level predictions for the Portuguese language at first shared task of the CASE @ ACL-IJCNLP 2022 workshop (Hürriyetoğlu, Ali and Mutlu, Osman and Duruşan, Fırat and Uca, Onur and Gürel, Alaeddin Selçuk et al., 2022), the Multilingual protest news detection subtask (Hürriyetoğlu et al., 2021a,b).

This article is organized as follows. In Section 2, reviews the related work. Section 3 details of subtask and data. Section 4 describes the methodology, while experiments results are discussed in Section 5. Section 6 brings the conclusions.

2 Related Work

Kalyan et al. (2021) proposed applying LSTM layers on top of 3 different models and combining the probabilities of each model in a soft voting manner. The models used were mBERT (Devlin et al., 2018), DistilmBERT (Sanh et al., 2019) and RoBERTa (Liu et al., 2019). They achieved a Macro F1 score of 0.7951 for the Portuguese.

Hettiarachchi et al. (2021) studied the use of long-range models such as big-bird and longformer as well as monolingual and multilingual models. They found that low-resource languages benefited from multilingual learning, but high-resource languages such as English will get better results from monolingual models. Their approach is similar to ours regarding the monolingual versus multilingual paradigm and their results demonstrated that multilingual models perfomance better than monolingual models in low data scenarios. Awasthy et al. (2021) work also agrees with the benefit from training with multilingual data on low-resource language cases.

Francesco Ignazio Re (2021) presented a disruptive perspective with the exploratory analysis of the dataset. Their conclusions approached differences in the use of state versus non-state conflict

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 184 - 188

actors based on conditional probabilities, and also identified an outlier in the English corpus via the Tf-Idf-weighted principal component analysis (PCA).

3 Subtask and Data

The dataset was provided by the organization of CASE 2022's first multilingual protest news detection shared task. Table 1 show some examples of dataset. The CASE 2022's a combination of CASE 2021 with new test data for Document classification subtask. These subtask focus on predicting whether a document contains information about some event related to protests. The dataset are composed of three languages: English, Spanish and Portuguese (Hürriyetoğlu et al., 2019a,b) for Socio-political Events in text domain. Table 2 shows the dataset distribution for each language. We random split the dataset in the ratio of 80% for the training and 20% validation set.

4 Methodology

We used pretrained transformer-based models for portuguese to investigate the classification performance of monolingual across multilingual models in scenario with low dataset resources. For this study, we selected two models and their multilingual versions:

- *BERT:* BERT (Devlin et al., 2018) is a pretrained language model trained using a masked language modeling and next sentence prediction objectives. The model has about 30k tokens in its vocabulary. Our version is the BERTimbau (Souza et al., 2020), trained on portuguese with the BRWAC dataset (Wagner Filho et al., 2018).

- *mBERT:* The multilingual cased version of BERT. It was trained on top of 104 languages using the wikipedia dataset. The training procedure was masked language modeling and next sentence prediction as in the original BERT, the main difference being the vocabulary size 110k tokens instead of 30k and the multilingual dataset.

- *RoBERTa:* The original RoBERTa (Liu et al., 2019) showed that increasing the vocabulary from around 30k to around 50k tokens and dropping the next sentence prediction training objective was beneficial for the model. Our version, trained on Portuguese, has a vocabulary size of 128k tokens and was trained on the Portuguese portion of OSCAR dataset and BRWAC dataset (Wagner Filho et al., 2018) for 100k steps.

- *xlm-RoBERTa:* the xlm-RoBERTa (Conneau et al., 2019) is a multilingual pretrained version of RoBERTa, which showed better performance than mBERT on NLI. It was pretrained similarly to Roberta but the training was done with 2.5TB of filtered CommonCrawl data containing 100 languages. The model has as vocabulary of about 250k tokens.

These models were optimized with a grid search optimization on held-out development set with a combination of finetuning hyperparameters provided by Table 3. We selected the best hyperparameter values based on 5 random seeds.

5 Results and Evaluation

All experiments were conducted on the Hugging's Face transformer library using one Nvidia A100 GPU (Choquette et al., 2021) for classify whether a document in Portuguese mentions an event or not. The models performance was evaluated by the macro F1-Scores on the validation set, which were created by splitting the dataset. The dataset for multilingual models was created by combining training data from each languages into one dataset. Table 5 shows the results of Portuguese document classification experiments on validation set using different sequence lengths and models. We can observe that increasing the max sequence length improves the performance on all tested models. Both multilingual versions of the models were better than their monolingual versions, showing that learning representations of other languages in the same task and domain can improve the model performance. The best result is shown in bold using the xlm-RoBERTa Large model achieving 0.8818 F1 score.

The results for the test set are shown in Table 4 with all models tested and the two best results submitted in the competition. According to the results, our best model became the best system for the document classification for Portuguese language. The xlm-RoBERTa model achieves the best result reaching 1st place with 0.8007 F1 score at Task 1 SubTask 1 Portuguese competition.

Table 1: Dataset examples for each language indicating the event mentions

Sentence	Language	Label
Publicidade Nessa propaganda dos		0
Explosão de carro-bomba deixa vários feridos em Israel Publi	Portuguese	1
Nos começos de 1964, instalara-se no cenário nacional a mesma divisão		0
OTHER STATES Kashmir unrest Protestors indulge in stone		1
Mass disconnection driv	English	0
403 Forbidden You don't have p		0
Las autoridades egipcias perdieron e		0
33 son los basquetbolistas argentino	Spanish	0
Un nuevo atentado sacudió al continente asiático. Do		1

Table 2: Dataset distribution

Language	Class 0	Class 1
Portuguese	1290	197
English	869	131
Spanish	869	131

Table 3: Hyperparameters for finetuning

Hyperparameter	CASE 2022
Max Epochs	{10, 20}
BatchSize	{8, 16, 32, 64}
Learning Rate	{2e-5, 3e-5, 4e-5, 5e-5}
Max Sequence Length	{128, 256, 512}
Learning Rate Decay	Linear
Warmup Ratio	0.1
Weight Decay	{0.1, 0.01}

6 Conclusion

In this paper, we have explored the capabilities of multilingual and monolingual language models on document classification of the CASE 2022 Task 1: Multilingual protest news detection. We demonstrate that multilingual transformer-based approach could be more competitive that monolingual transformer-based model in scenarios that have low data resources of a specific language and more data of other languages can help achieve a best performance. The proposed xlm-RoBERTa model achieved the 1st place for the Portuguese language with 0.8007 F1 Score on Test set.

These results illustrate the importance of increasing the maximum sequence length for document classification. As future work, it would be interesting to extend the study to architectures with much longer input sequences. We also investigate other methodologies based on ensemble approach, data augmentation and few shot models.

Limitations

Recent works demonstrated that monolingual language models achieves better performance than multilingual models in NLP downstream task. The dataset size for a specific language task can be an issue (scenarios with low amount of data resource). Our experimental results demonstrate that using a multilingual model with more data from other languages achieves a better result than a monolingual model trained only in a specific language. The low amount of data for non-English language be a difficult for training monolingual language models. Finally, in this case, the size of maximum sequence length has a big impact in performance and transformers-based models size resulting a requirement of large GPU resources to processing long texts.

Ethics Statement

Most of the recent work on language models rely on vast amount of unannotated data to achieve good results, which means that these models are very likely to be training on harmful content to some degree. It is possible that the bias present in the pretraining continues to play a role after the fine-tuning of the model. The amount of bias influencing the model is yet to be quantified and future work should try to measure this before and after fine-tuning on specific data.

Table 4: Document classification results for Portuguese test data set. Best results is in Bold.

Model	Training data	Macro F1
team1	-	0.7985
team2	-	0.7922
BERT	pt	0.7372
mBERT	pt + en + es	0.7525
RoBERTa	pt	0.7732
xlm-RoBERTa	pt + en + es	**0.8007**

Table 5: Macro F1 results of document classification experiments for Portuguese using different sequence lengths and models on dev set. Best results is in Bold.

Model	Training data	Seq. Length	Accuracy	Macro F1
BERT	pt	128	0.9142	0.8443
		256	0.9199	0.8491
		512	0.9261	0.8533
mBERT	pt + en + es	128	0.9076	0.8528
		256	0.9086	0.8542
		512	0.9136	0.8600
RoBERTa	pt	128	**0.9328**	0.8641
		256	0.9327	0.8696
		512	0.9362	0.8721
xlm-RoBERTa	pt + en + es	128	0.9246	0.8727
		256	0.9293	0.8781
		512	0.9310	0.**8818**

Acknowledgements

Authors thank Center of Excellence in Artificial Intelligence (CEIA) of Federal University of Goiás for their support in conducting this study.

References

Parul Awasthy, Jian Ni, Ken Barker, and Radu Florian. 2021. Ibm mnlp ie at case 2021 task 1: Multigranular and multilingual event detection on protest news. pages 138–146.

Jack Choquette, Wishwesh Gandhi, Olivier Giroux, Nick Stam, and Ronny Krashinsky. 2021. Nvidia a100 tensor core gpu: Performance and innovation. *IEEE Micro*, 41(2):29–35.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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.

Fırat Duruşan, Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Çağrı Yoltar, Burak Gürel, and Alvaro Comin. 2022. Global contentious politics database (glocon) annotation manuals.

Dennis Atzenhofer Niklas Stoehr Francesco Ignazio Re, Daniel Véegh. 2021. Team "dadefrni" at case 2021 task 1: Document and sentence classification for protest event detection. pages 138–146.

Hansi Hettiarachchi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed Gaber. 2021. Daai at case 2021 task 1: Transformer-based multilingual socio-political and crisis event detection. pages 120–130.

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, Erdem Yörük, Osman Mutlu, Deniz Yüret, and Aline Villavicencio. 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, Hristo Tanev, Vanni Zavarella, Reyyan Yeniterzi, Osman Mutlu, and Erdem Yörük. 2022. 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).

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.

Hürriyetoğlu, Ali and Mutlu, Osman and Duruşan, Fırat and Uca, Onur and Gürel, Alaeddin Selçuk, Benjamin Radford, Yaoyao Dai, Hansi Hettiarachchi, Niklas Stoehr, Tadashi Nomoto, Milena Slavcheva, Francielle Vargas, Aaqib Javid, Fatih Beyhan, and Erdem Yörük. 2022. 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).

Pawan Kalyan, Duddukunta Reddy, Adeep Hande, Ruba Priyadharshini, Sakuntharaj Ratnasingam, and Bharathi Chakravarthi. 2021. Iiitt at case 2021 task 1: Leveraging pretrained language models for multilingual protest detection. pages 98–104.

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. *CoRR*, abs/1907.11692.

Robin Tamarelli McKeon and Drew H. Gitomer. 2019. Social media, political mobilization, and high-stakes testing. *Frontiers in Education*, 0. [Online; accessed 2022-09-25].

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

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)*.

Jorge A Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brwac corpus: A new open resource for brazilian portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.