ComMA@FIRE 2020: Exploring Multilingual Joint Training across different Classification Tasks

Ritesh Kumar[a,b], Bornini Lahiri[c], Atul Kr. Ojha[e,d] and Akanksha Bansal[d]

[a]*Department of Linguistics, K.M. Institute of Hindi and Linguistics, Dr. Bhimrao Ambedkar University, Agra*
[b]*Centre for Transdisciplinary Studies, Dr. Bhimrao Ambedkar University, Agra*
[c]*Indian Institute of Technology, Kharagpur*
[e]*DSI, NUIG, Galway*
[d]*Panlingua Language Processing LLP, New Delhi*
[d]*Panlingua Language Processing LLP, New Delhi*

Abstract

In this paper, we give a description of the systems submitted to the three tracks of FIRE 2020 - Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC), Sentiment Analysis of Dravidian Languages in Code-Mixed Text and Event Detection from News in Indian Languages (EDNIL). While the first two tasks were binary and multi-class text classification problems, EDNIL was a sequence classification problem. For all the three tracks, we jointly fine-tuned mBERT, DistilBERT, RoBERTa and XLM-R using the dataset from all the languages for the given task.

Keywords

LaTeX mBERT, XLM-R, DistilBERT, Event Detection, Offensive Language, Sentiment Analysis, Multilingual Systems

1. Introduction

In the last couple of years, Transformers-based models such as BERT and RoBERTa have proved to be quite successful for a variety of NLP tasks including a range of text classification tasks. One of the most promising aspects of these models is the fact that they could be pre-tuned on huge amount of raw texts and later fine-tuned on downstream tasks with relatively fewer train instances. While this approach has proved to be quite successful with resource-rich languages such as English, there are two issues with low-resource languages -

- There are only a handful of languages for which the pre-trained models are available. Since pre-training is quite expensive - both in terms of computational resources and infrastructure required as well as linguistic resources needed for building reasonably good models - these pre-trained models are not available for a large number of languages.
- For a large number of tasks, sufficient data for fine-tuning the model is also not available in a large number of languages.

Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad
ritesh78_llh@jnu.ac.in (R. Kumar); bornini@hss.iitkgp.ac.in (B. Lahiri); panlingua@outlook.com (A. Bansal)
0000-0002-5151-2546 (R. Kumar); 0000-0001-7841-3292 (B. Lahiri); 0000-0002-9800-9833 (A.Kr. Ojha)

© 2020 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

The lack of sufficient resources for most of the world's languages has been a classic problem in NLP, resulting in a hugely imbalanced progress in the development of language technologies for different languages (see [1] for an excellent analysis of this issue in NLP). One of the ways of handling this problem is to make use of multilingual and transfer learning methods for training the systems. In order to exploit these methods, pre-trained models such as mBERT (multilingual BERT) and XLM-R (cross-lingual model of RoBERTa) have been made available.

In this paper, we discuss our experiments using multilingual joint training methods for three different kinds of tasks - sentiment analysis, offensive language identification and event detection - and report their results. All these three tasks were organised as shared tasks under the aegis of the Forum of Information Retrieval Evaluation 2020 (FIRE 2020).

2. The Tasks and Datasets

We participated in the following tasks at FIRE 2020.

- **Sentiment Analysis of Dravidian Languages in Code-Mixed Text [2]:** This task was a message-level polarity classification task. The dataset consisted of YouTube comments in Tamil and Malayalam (see Table 1) [3, 4]. Each comment was annotated with one of the following categories - positive, negative, neutral and mixed emotions. In addition to this, the task also involved a language identification task where the comments not in the intended language were also to be identified.
- **Event Detection from News in Indian Languages (EDNIL)[5]:** This task involved identifying those segments of text which contained an event within a news article. It involved two sub-tasks - in sub-task 1, man-made and natural disasters were to be automatically identified; in sub-task 2, sub-types of man-made disasters (16 sub-types) and natural disasters (21 sub-types), casualties, time, place and reason of the disasters were to be identified. The dataset was provided in 5 Indian languages - Hindi, Bengali, Marathi, Tamil and English. We submitted our system for only sub-task 1 and 3 languages - Hindi, Bengali and English (see Table 2).
- **Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) [6]:** This task consisted of two sub-tasks. In sub-task A, the data was annotated as HOF (Hate Speech and Offensive Language) and NOT (Not Offensive). In sub-task B, HOF instances were further classified into Hate Speech, Offensive and Profane. The dataset was provided in 3 languages - Hindi, English and German (see Table 3). In this task, we participated in both the sub-tasks in all the three languages.

3. Experiments and Results

For each of the language in each of the task, we fine-tuned the following models -

- cased multilingual BERT

Table 1
Dravidian Language Sentiment Analysis Dataset

		TOTAL	Positive	Negative	Mixed	Unknown	Other_Language
Malayalam	**Train**	4830	2015	549	289	1341	636
	Dev	537	222	51	333	1501	696
	Test	1348	565	138	70	398	177
Tamil	**Train**	11283	7579	1447	1281	606	368
	Dev	1254	855	164	139	67	29
	Test	3149	2075	424	377	173	100

Table 2
EDNIL Dataset

		TOTAL	Man-made Disaster	Natural Disaster
Bangla	**Train**	4095	3386	709
	Dev	1025	847	178
Hindi	**Train**	3571	1748	1823
	Dev	893	437	456
English	**Train**	3881	3019	862
	Dev	971	755	216

Table 3
HASOC Dataset

Language		Sub-task A			Sub-task B			
		TOTAL	NOT	HOF	TOTAL	HATE	OFFN	PRFN
Hindi	**Train**	2599	1851	748	748	214	403	131
	Dev	354	265	99	99	20	62	17
	Test	663	466	197	197	56	87	27
English	**Train**	3300	1653	1647	1647	143	285	1219
	Dev	408	199	209	209	15	36	158
	Test	814	391	423	423	25	82	293
German	**Train**	2000	1431	569	569	146	140	387
	Dev	373	269	104	104	29	24	51
	Test	526	392	134	134	24	36	88

- cased multilingual DistilBERT
- XLM-R (cross-lingual, multilingual model of RoBERTa)

Two sets of models were trained for each of these - one was a separate model for each language and the other was a joint multilingual model using the dataset of all the languages available for each task. Thus for the sentiment analysis task, the multilingual model was trained using Malayalam and Tamil datasets, for EDNIL task, Bangla, Hindi and English datasets were

Table 4
F-score of different models in different tasks

Task	Language	mBERT	mBERT(j)	mDistilBERT	mDistilBERT(j)	RoBERTa	XLM-R
EDNIL	BEN	–	0.356	–	0.369	–	0.385
EDNIL	HIN	–	0.493	–	0.504	–	0.510
EDNIL	ENG	–	0.585	–	0.586	–	0.587
Senti	MAL	0.51	0.52	0.52	0.52	0.53	–
Senti	TAM	0.58	0.58	0.56	0.57	0.58	–
HASOC A	HIN	0.72	0.72	0.70	0.71	0.58	0.58
HASOC A	ENG	0.86	0.87	0.87	0.87	0.87	0.83
HASOC A	DE	0.82	0.84	0.84	0.82	0.85	0.78
HASOC B	HIN	0.69	0.70	0.68	0.70	0.63	0.63
HASOC B	ENG	0.73	0.79	0.78	0.81	0.78	0.74
HASOC B	DE	0.76	0.75	0.76	0.76	0.77	0.72

used and for HASOC task, Hindi, English and German datasets were used.

The training was carried out using the Simple Transformers library. We used a batch size of 12, maximum sequence length of 512 and a learning rate of 6e-5 for fine-tuning each of the models. The models were trained for 15 epochs. For all the other hyperparameters, the default setting of the library was retained. The results of each of the model for each of the task is summarised in Table 4. In this table we report the weighted F-score of all the systems that we submitted for different tasks.

In addition to this, the Dravidian sentiment analysis task involved an additional language classification task. We used an SVM classifier with character 5-gram and word unigram as features for this task. We experimented with different combinations of character bigram to 5-gram and word unigram to trigram feature set to find the optimum features for the language identification. In this task, the first step was that of language identification. If the language was either Tamil or Malayalam then the test instance was classified for its sentiment polarity. The scores reported here is the final score of the classification pipeline and includes errors made by the language identification system.

For EDNIL, we did not submit the models trained on individual languages because there was no significant difference between the multilingual and the monolingual models. One reason for this apparent failure of joint multilingual training could be the kind of data released for EDNIL. The dataset is taken from the newspapers, which are heavily edited (and so kind of not very natural) and removes all the characteristics of a natural multilingual communication such as borrowings and code-mixing. Moreover the three languages that we jointly trained on - Hindi, Bangla and English - all three did not share the script. Hence it is quite natural that knowledge (or more appropriately patterns) from one language could not be transferred to the other languages, hence, no improvement is noticed with joint training.

However, in case of Dravidian sentiment analysis tasks, the dataset was taken from social media which are more naturalistic and depicts the properties of multilingual communication such as code-mixing and switching as well as the use of the same script (Roman) for writing all

languages including Malayalam and Tamil. As such we witness knowledge transfer from one language to the other in case of joint multilingual training in these tasks.

For HASOC, the complete test set is not yet released and the organisers have also not released the scores for each model submitted to the task separately. However, only 15% of the test set was supposed to be unseen - the labels for rest of the instances used for testing has been released by the organisers. So the scores reported here are those obtained on the released test set (and so do not exactly match those on the leaderboard since those scores are based on the additional test set).

In general, we see that the joint training across different tasks and languages have yielded marginal improvements in performance. One of the reasons could be the small dataset size for all the languages. Generally in cases of multilingual training at least one of the languages have a relatively large dataset and the knowledge from that dataset is transferred to other languages. However, because of lack of this, the improvement is marginal but nevertheless joint training seems to give some improvement.

4. Conclusion

In this paper, we have presented a description of our system submitted to three tracks at FIRE 2020 - Sentiment Analysis for Dravidian Languages, EDNIL and HASOC. We have explored multilingual joint training of three transformers-based models - mBERT, mDistilBERT and XLM-R - and evaluated their performance vis-a-vis monolingual models in each of the tasks. The results demonstrate that multilingual joint training provides significant advantage in processing what has been traditionally termed 'difficult' (or noisy) datasets - the naturalistic, multilingual, multi-scriptal social media conversations - while there is no significant gain when the datasets are from what has been traditionally termed 'clean', carefully edited datasets from newspapers (in case of EDNIL). Of course, since these results are obtained on different tasks, in order to validate these results, they also need to be tested on the same task with different kinds of datasets (or even better with the two versions of the same dataset). However, notwithstanding this extra variable, it is quite apparent that multilingual methods are quite effective for classification and processing of multilingual documents while not so for the monolingual documents.

References

[1] P. Joshi, S. Santy, A. Budhiraja, K. Bali, M. Choudhury, The state and fate of linguistic diversity and inclusion in the NLP world, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6282–6293. URL: https://www.aclweb.org/anthology/2020.acl-main.560. doi:10.18653/v1/2020.acl-main.560.

[2] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, E. Sherly, J. P. McCrae, Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text, CEUR Workshop Proceedings, CEUR-WS.org, 2020.

[3] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, J. P. McCrae, Corpus creation for sentiment analysis in code-mixed Tamil-English text, in: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), European Language Resources association, Marseille, France, 2020, pp. 202–210. URL: https://www.aclweb.org/anthology/2020.sltu-1.28.

[4] B. R. Chakravarthi, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, A sentiment analysis dataset for code-mixed Malayalam-English, in: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), European Language Resources association, Marseille, France, 2020, pp. 177–184. URL: https://www.aclweb.org/anthology/2020.sltu-1.25.

[5] B. Dave, S. Gangopadhyay, P. Majumder, P. Bhattacharya, S. Sarkar, S. L. Devi, Overview of the FIRE 2020 EDNIL track: Event Detection from News in Indian Languages, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, CEUR Workshop Proceedings, CEUR-WS.org, 2020.

[6] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Multiple Languages), in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, CEUR, 2020. URL: http://ceur-ws.org/.