Applying Multilingual and Monolingual Transformer-Based Models for Dialect Identification

Cristian Popa and Vlad Ștefănescu

Classification of Romanian and Moldavian dialects by training and validating on the MOROCO[1] dataset and testing on tweets.

Our work involves a comparison of transformer-based models, as suggested by the title, both monolingual and multilingual:

- mBERT[2]
- XLM[3]
- XLM-RoBERTa[4]
- Cased and uncased Romanian BERT[5]

Of these, the first 4 are multilingual, trained on large corpora of at least 100 different languages, while the last ones are trained from scratch on Romanian text data.

The metrics we analyze in our experiments are the AUC and macro-F1:

Table 1: Qualitative results of all the models used.

Models	News Extracts		Tweets	
	AUC	Macro-F1	AUC	Macro-F1
mBERT	0.9915	0.9607	0.7744	0.6979
XLM	0.9839	0.9438	0.7899	0.7113
XLM-R	0.9944	0.9694	0.7916	0.7227
Cased Rom. BERT	**0.9955**	0.9729	0.8386	0.7460
Uncased Rom. BERT	0.9948	0.9724	**0.8549**	0.7627
SVM Ensemble	0.9868	**0.9752**	0.8428	**0.7752**

Our experiments show that the **monolingual models outclass the multilingual ones** on both the MOROCO (news extracts) and tweets datasets.

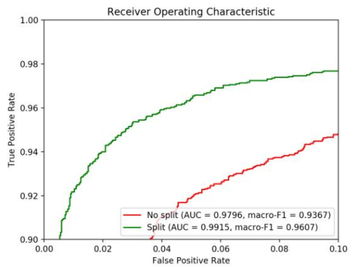
We build a final **SVM ensemble** from the predictions of all the previous models, that achieves the highest macro-F1 score at the optimal threshold.

Our SVM model manages to secure the **second place** in the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2020), where we participated under team "Anumiți":

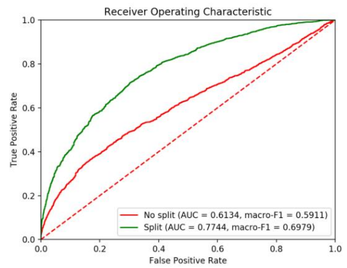
Table 2: Best macro-F1 scores for all the submissions on the RDI shared task.

Models	Macro-F1
Tubingen	0.787592
Anumiți (SVM Ensemble)	**0.775178**
Phlyers	0.666090
SUKI	0.658437
UPB	0.647577
UAIC	0.555044
akanksha	0.481325
The Linguistadors	0.429412

A big reason for this performance is due to the tokenization we perform on the input datasets, by **splitting samples into sentences** using the spaCy[6] package:



(a) Performance of mBERT on news extracts.

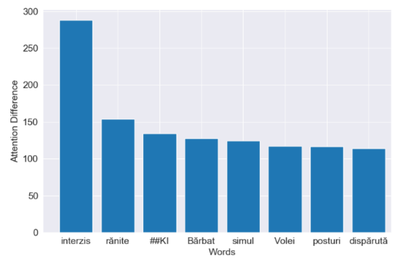


(b) Performance of mBERT on tweets.

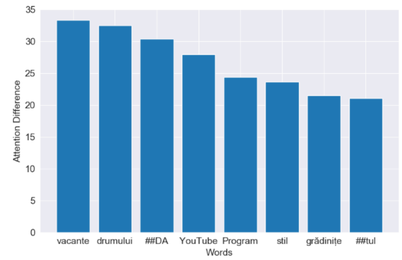
We additionally perform an **analysis on the attention** of the Romanian BERT models, since these achieved the best results.

The analysis is done starting from open-source code[7], which we modified to find **"differentiating tokens"** between the two dialects. We defined differentiating tokens as those that the model pays a lot of attention to when predicting one dialect and a lower amount of attention when predicting the other. Top differentiating tokens maximize the difference in paid attention.

Knowing this, we showcase the differentiating tokens of both dialects for the cased version of Romanian BERT:

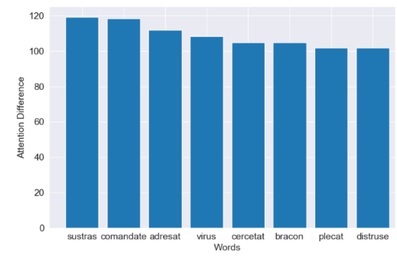


(a) Differentiating tokens for the Romanian dialect.

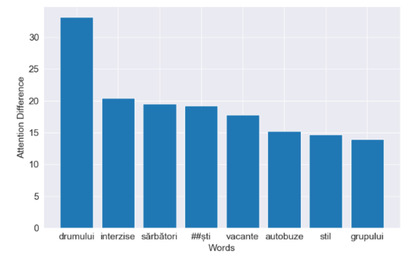


(b) Differentiating tokens for the Moldavian dialect.

As well as for the uncased version of Romanian BERT:



(a) Differentiating tokens for the Romanian dialect.



(b) Differentiating tokens for the Moldavian dialect.

We open-sourced our code, involving pre-processing, fine-tuning of the transformer-based models and analyzing the results:

https://github.com/CristianViorelPopa/transformers-dialect-identification

References

[1] Andrei Butnaru and Radu Tudor Ionescu. 2019. MOROCO: The Moldavian and Romanian dialectal corpus
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding
[3] Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining
[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzm´an, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale
[5] Ștefan Daniel Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT
[6]Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing
[7] https://github.com/clarkkev/attention-analysis