Dialect Identification through Adversarial Learning and Knowledge Distillation on Romanian BERT

George-Eduard Zaharia[1], **Andrei-Marius Avram**[1, 2],
Dumitru-Clementin Cercel[1], **Traian Rebedea**[1]
University Politehnica of Bucharest, Faculty of Automatic Control and Computers[1]
Research Institute for Artificial Intelligence, Romanian Academy[2]
{george.zaharia0806, andrei_marius.avram}@stud.acs.upb.ro
{dumitru.cercel, traian.rebedea}@upb.ro

Abstract

Dialect identification is a task with applicability in a vast array of domains, ranging from automatic speech recognition to opinion mining. This work presents our architectures used for the VarDial 2021 Romanian Dialect Identification subtask. We introduced a series of solutions based on Romanian or multilingual Transformers, as well as adversarial training techniques. At the same time, we experimented with a knowledge distillation tool in order to check whether a smaller model can maintain the performance of our best approach. Our best solution managed to obtain a weighted F1-score of 0.7324, allowing us to obtain the 2nd place on the leaderboard.

1 Introduction

Dialect identification has attracted researchers from both the fields of speech and natural language processing, because of its wide appliance in different tasks such as automatic speech recognition (Biadsy, 2011), machine translation (Salloum et al., 2014), or opinion mining (Salamah and Elkhlifi, 2014). VarDial (Zampieri et al., 2020) is an yearly workshop that deals with boosting the research in this direction by creating computational resources for languages that are closely related with each other, varieties in language, and dialects. This year's edition (Chakravarthi et al., 2021) was composed of four subtasks: (1) Dravidian Language Identification (DLI), (2) Romanian Dialect Identification (RDI), (3) Social Media Variety Geolocation (SMG), and (4) Uralic Language Identification (ULI). We chose to participate in the second subtask of the workshop, the RDI subtask. This subtask was also proposed in the previous edition of the workshop (Gaman et al., 2020), but this time the participants are given an augmented version of the Moldavian and Romanian Dialectal Corpus (MOROCO) dataset (Butnaru and Ionescu, 2019) that contains texts from the news domain. Also, as in the previous edition, the test set comes from another domain, so cross-domain algorithms must be employed in order to maximize the results.

Romanian language resources are in an ongoing process of maturity, and the language is slowly starting to gain the datasets necessary to not be considered under-resourced anymore. Some of the recent publicly available Romanian corpora include the Large Romanian Sentiment DataSet (LaRoSeDa) (Tache et al., 2021), the Romanian Named Entity Corpus (RONEC) (Dumitrescu and Avram, 2020), and the Romanian version of the Cross-lingual Question Answering Dataset (xQuAD) (Artetxe et al., 2020). Moreover, with the rise of Transformer-based pretrained language models (Vaswani et al., 2017), some Romanian model versions have also been created (Dumitrescu et al., 2020; Masala et al., 2020). The speech resources are also starting to catch-up with the introduction of the Romanian Speech Corpus (RSC) (Georgescu et al., 2020) which was recently released for public usage, counting around 100 hours of speech, and with the introduction of a deep neural network architecture based on DeepSpeech2 (Amodei et al., 2016) for automatic speech recognition (Avram et al., 2020b).

In this work, we proposed a series of models based on Transformers, pre-trained on the Romanian language, and aimed to tackle the dialect identification task. By using different techniques, including adversarial training (Goodfellow et al., 2014b), knowledge distillation (Hinton et al., 2015), or Generative Adversarial Networks (GANs) (Goodfellow et al., 2014a), we managed to obtain good scores, allowing us to classify 2nd in the RDI subtask organized at VarDial 2021.

The current work is structured as follows. The next section presents the state of the art regarding the Romanian-Moldavian dialect identification

Proceedings of the 8th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 113–119
April 20, 2021 ©2021 Association for Computational Linguistics

task. Section 3 outlines the methods created by us in order to tackle the previously mentioned challenge, while section 4 describes the results and displays the error analysis we conducted. Section 5 concludes the work and features some future improvements that can be made to further increase the performance.

2 Related Work

The third edition of the VarDial evaluation campaign (Zampieri et al., 2019) took place in 2019 and presented another challenging task for the Romanian language - Moldavian vs. Romanian Cross-dialect Topic identification. The task was composed of three smaller subtasks in which the participants had to (1) discriminate between the Moldavian (MD) and the Romanian (RO) dialects, (2) use Moldavian samples to classify Romanian samples by topic (MD → RO), and (3) use Romanian samples to classify Moldavian samples by topic (RO → MD). The highest macro F1-score on the first subtask was 89.50%, achieved by the `DTeam` team (Tudoreanu, 2019) with an ensemble model that combines a skip-gram convolutional neural network (CNN) (Kim, 2014) using the softmax loss and a CNN that was trained using a triplet loss. The highest scores for the cross-dialect subtasks 2 and 3 were obtained by the `tearsofjoy` team (Wu et al., 2019) that used a linear Support Vector Machine (SVM) classifier trained on a combination of character and word n-gram features. They obtained a 61.15% F1-score (macro) for the MD → RO subtasks and a 55.33% F1-score (macro) for the RO → MD subtask. Onose et al. (2019) adopted a non-Transformer approach and employed the usage of neural network models with Bidirectional Long Short-Term Memory cells (Hochreiter and Schmidhuber, 1997), Bidirectional Gated Recurrent units (Cho et al., 2014), as well as a Hierarchical Attention Network (Yang et al., 2016).

The fourth edition of VarDial (Gaman et al., 2020) occured in 2020 and came with the RDI task, a binary classification task where participants had to identify the dialect of a given text - either Romanian or Moldavian. To differentiate from the previous edition, the evaluation set was taken from another domain, namely Twitter messages. The highest F1-score (macro) of 78.75% was achieved by the `Tubingen` team (Çöltekin, 2020) by using an ensemble of SVMs trained on word and character n-grams. Also, this edition saw the appliance of the Romanian Bidirectional Encoder Representations from Transformers (BERT) by two teams (Popa and Ștefănescu, 2020; Zaharia et al., 2020a), with the best performing variant obtaining a 77.51% F1-score (macro).

3 Method

3.1 Transformer-based Models

Our architectures are based on multiple Transformer-based models, considering their performance on various natural language processing (NLP) tasks (Avram et al., 2020a; Dima et al., 2020; Ionescu et al., 2020; Paraschiv et al., 2020; Paraschiv and Cercel, 2019; Tanase et al., 2020b,a; Vlad et al., 2020; Zaharia et al., 2020b). For Romanian Dialect Identification, we employed the usage of a Transformer model extensively pre-trained on the Romanian language (Dumitrescu et al., 2020), as well as two multilingual models, namely XLM-RoBERTa (Conneau et al., 2019) and multilingual BERT (mBERT) (Pires et al., 2019). Their performance on the Romanian language is lower when compared to Romanian BERT, however, in an ensemble, their predictions can prove to increase the score of our approach.

3.2 Generative Adversarial Network Applied on Romanian BERT

Moreover, using a training technique similar to the ones employed by GANs proved to improve the performance of several NLP models (Croce et al., 2020). Similarly, we augment our Romanian BERT architecture with a generator, as well as a discriminator. The generator receives as input a 100-dimensional noise vector and produces an output vector as similar to real inputs as possible. Moreover, the discriminator acts as a classifier but, instead of only being forced to distinguish between the two classes of the RDI task (i.e., *RO* or *MD*), it also has the purpose to identify whether the input is fake or not and classify it accordingly, into a third class. The discriminator is penalized if it classifies a fake input as a true one or vice versa. After the training process, the generator components and the third output of the discriminator are disabled, therefore our architecture works as a classifier based on Romanian BERT features.

3.3 Knowledge Distillation Applied on Romanian BERT

We continued our experiments by applying a knowledge distillation technique, in order to check whether the high performance of Romanian BERT is maintained after it is distilled into a smaller model. For this approach, we used TextBrewer (Yang et al., 2020), a tool that receives as input a teacher model and trains a student model such that the latter is able to closely replicate the behavior of the former. We used Romanian BERT as the teacher, a Transformer-based model with 12 hidden layers, thus implying a large number of parameters. Moreover, the student model is also based on Transformers, however, it is not pre-trained on any corpus and it has only 3 hidden layers, instead of 12. This reduction greatly decreases the computational resources required for further fine-tuning or prediction.

3.4 Adversarial Training

Adversarial training has the purpose of enhancing the robustness of the model and, as a consequence, increase its performance in certain scenarios (Karimi et al., 2020). The system works by introducing adversarial perturbations at the level of the Transformer embeddings. The process turns into a minimization problem, with the purpose of determining the worst perturbations while minimizing the loss function.

The following formulas present the process of obtaining the adversarial perturbations, based on the gradient of the loss function g:

$$g = \nabla_x log\, p(y|x; \hat{\theta}) \tag{1}$$

$$r_{adv} = -\epsilon \frac{g}{||g||_2} \tag{2}$$

where $\hat{\theta}$ is a copy of the model's parameters, r_{adv} are the perturbations, and ϵ is the dimension of the perturbations.

After computing the perturbations, they are then added to the Transformer embeddings and an adversarial loss is obtained, given by Eq. 3:

$$-log\, p(y|x + r_{adv}; \theta) \tag{3}$$

The final loss represents the sum between the adversarial loss and the simple loss obtained by passing the unaltered input through the neural network.

3.5 Custom Selection Technique

The most important element that influenced the performance of our models is represented by the way we selected the training entries, as well as how we established the threshold for which an entry can be considered Romanian or Moldavian.

Firstly, the training dataset contains long entries, most of them surpassing the 512-token limit input by our Transformer models. At the same time, the validation entries are much shorter, with an average length of just 20 words.

Therefore, the first step we performed was to split the training entries into sentences and label each sentence with the label of the initial entry. However, the number of training entries was greatly increased, from 39,487 to 431,875. Considering this large number, we decided to filter the training entries, inasmuch as only the most relevant ones were kept for the final fine-tuning process. To do this, we initially trained our architecture, based on the Romanian BERT model, on the original validation entries. After four epochs, we tested the model on the split training entries and we selected only the ones that predicted Romanian or Moldavian with the confidence of over 95%. This way, we were able to select only the entries that are the closest in structure and context to the one from the validation dataset and, presumably, from the test one. We reduced the number of the split training entries to 158,363.

For determining the prediction threshold, we trained our architectures on the previously mentioned entries and, after the final epoch, we discovered a prediction threshold that maximized the performance. We performed the selection by trying different values such as, if the confidence of a prediction surpasses the threshold, then it is classified as Moldavian, for example, if not, it is classified as Romanian. The optimal value of the threshold was 0.21.

3.6 Machine Learning Approaches

We also experimented with various machine learning techniques, such as SVMs, Random Forest, Multinomial Naive Bayes, and Logistic Regression, alongside character n-gram features.

4 Experiments

4.1 Dataset Analysis and Preprocessing

The dataset is the one provided for the RDI subtask of the VarDial 2021 competition. There are three subsets, one for training, one for validation, and one for testing. The validation and testing datasets (5,237 and 5,282 entries, respectively) contain very short entries, with an average length of 20 words. At the same time, the training dataset (39,487 entries) contains much longer texts, many of them surpassing 512 words. The class distribution is relatively balanced, with 21,366 Romanian entries and 18,121 Moldavian entries in the training dataset plus 2,625 and 2,612 entries, respectively, in the validation one.

In terms of preprocessing, we standardized the punctuation, by removing repeated characters such as question marks. Moreover, we cleaned the entries that contained an unnecessary number of whitespaces.

4.2 Implementation Details

For running the Transformer-based models we used Adam with weight decay optimizer (AdamW) (Kingma and Ba, 2014) and an epsilon value of *1e-8*. We fine-tuned them for four epochs with a learning rate of *2e-5*. Moreover, for the GAN approach, we allowed the generator to backpropagate its loss every 200 steps, such that the discriminator, which also performs as the classifier, had a small advantage.

4.3 Results

Table 1 presents all the results obtained with neural network approaches. The best results are obtained by the ensemble used by taking the predictions from the six deep learning models used in our experiments. With a weighted F1-score of 0.7324, the ensemble slightly surpasses the Romanian BERT model trained under adversarial circumstances, which scored a weighted F1 of 0.7318. The small performance improvement is noticeable on the validation dataset, as well, the ensemble scoring a 0.8564 weighted F1.

The adversarial training technique helps the Romanian BERT model improve its performance, considering the higher weighted F1-score on the validation dataset, 0.8559, obtained by the adversarial model when compared to the standard counterpart (0.8492). In contrast, the GAN training approach does not help the model achieve improved performance. The weighted F1-score is slightly higher than the standard Romanian BERT. However, when compared to the adversarial training technique, GAN+Romanian BERT lacks behind in terms of performance.

The distilled model obtained by using TextBrewer with Romanian BERT comes last in terms of performance in the group of the Romanian BERT-based models. The weighted F1-scores of 0.7891 and 0.6744 obtained on the validation and test datasets are behind all the scores obtained by using variations of the Romanian BERT. The lack of performance can be attributed to the lower number of parameters of the distilled model, which it is not able to grasp the distinctive features of Romanian and Moldavian entries as well as the full model.

The last models in terms of performance are Multilingual BERT and XLM-RoBERTa. Even though the pre-training corpus of XLM-RoBERTa is bigger, the model is surpassed by mBERT. One reason for this can be represented by the inclusion of more Romanian entries in the mBERT pre-training corpus.

Table 2 contains the results obtained by using various machine learning techniques alongside character n-gram features. The scores are much lower when compared to the ones achieved by the neural network approaches. The best performing model is the SVM trained with the parameter *C* equal to 3. The model achieves a weighted F1-score of 0.6298, 0.1324 lower than the worst score obtained by the neural network approaches.

4.4 Error Analysis

Most misclassifications come from the inability of the models to identify dialect-specific features in the input entries. Taking as an example the validation dataset, many entries that are classified as Romanian or Moldavian have no surface differences in terms of structure or the used words (i.e., ”*Ultimele știri despre coronavirus NE*” is labeled with the Romanian dialect, while ”*Cum te protejezi împotriva coronavirusului NE*” is Moldavian). At the same time, the entries that are classified with high confidence as either Romanian or Moldavian are the ones with unmasked named entities, that are also present in the training dataset (i.e., *”Arafat”*, *”Ceban”*, *”Igor”*) or dialect-specific words (i.e., *”raional”*).

Some entires are obstructed by the masked named entities (i.e., ”*Este criză în NE și NE NE*”, ”*FOTO-VIDEO. Ministrul NE NE NE și NE NE NE la NE NE Nu este vorba*”) and therefore our models cannot properly identify features specific to one dialect or the other.

Table 1: Deep learning results.

Model	Validation			Test		
	macro-F1	weighted-F1	micro-F1	macro-F1	weighted-F1	micro-F1
Romanian BERT	0.8492	0.8492	0.8495	-	-	-
Romanian BERT + Adversarial Training	0.8558	0.8559	0.8564	0.7319	0.7318	0.7319
Multilingual BERT	0.8097	0.8097	0.8098	-	-	-
XLM-RoBERTa	0.7619	0.7619	0.7622	-	-	-
Romanian BERT + GAN	0.8516	0.8516	0.8523	-	-	-
Romanian BERT + TextBrewer	0.7891	0.7891	0.7893	0.6743	0.6744	0.6749
Ensemble	**0.8564**	**0.8564**	**0.8566**	**0.7324**	**0.7324**	**0.7324**

Table 2: Machine learning results on the validation dataset.

Model	Validation		
	macro-F1	weighted-F1	micro-F1
SVM	**0.6297**	**0.6298**	**0.6324**
Random Forest	0.5216	0.5218	0.5379
Multinomial Naive Bayes	0.5726	0.5728	0.5921
Logistic Regression	0.6250	0.6251	0.6270

Moreover, the performance difference between the validation and test sets (i.e., 0.8564 vs. 0.7324) can be attributed to the selection technique we used for filtering the training entries. The new inputs are chosen such that they are similar to the validation entries, not the test ones.

5 Conclusions and Future Work

This work presents our approaches for the Romanian Dialect Identification subtask organized by VarDial 2021. We proposed a series of systems based on state-of-the-art, Transformer-based models, which imply the usage of adversarial techniques for improving the robustness. At the same time, we also experimented with TextBrewer, a knowledge distillation tool that allows us to compress a teacher model into a student model such that the performance can be maintained while reducing the size. Moreover, by using an ensemble of all the models we experimented with, we managed to improve the overall performance.

For future work, we intend to experiment with different variants of adversarial training for increasing the scores obtained by our models.

References

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Andrei-Marius Avram, Dumitru-Clementin Cercel, and Costin-Gabriel Chiru. 2020a. Upb at semeval-2020 task 6: Pretrained language models for definitionextraction. *arXiv preprint arXiv:2009.05603*.

Andrei-Marius Avram, PĂIȘ Vasile, and Dan Tufis. 2020b. Towards a romanian end-to-end automatic speech recognition based on deepspeech2. In *Proc. Rom. Acad. Ser. A*, volume 21, pages 395–402.

Fadi Biadsy. 2011. *Automatic dialect and accent recognition and its application to speech recognition*. Ph.D. thesis, Columbia University.

Andrei Butnaru and Radu Tudor Ionescu. 2019. Moroco: The moldavian and romanian dialectal corpus. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 688–698.

Bharathi Raja Chakravarthi, Mihaela Găman, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial Evaluation Campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger

Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Çağrı Çöltekin. 2020. Dialect identification under domain shift: Experiments with discriminating romanian and moldavian. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 186–192.

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. *arXiv preprint arXiv:1911.02116*.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. Gan-bert: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119.

George-Andrei Dima, Andrei-Marius Avram, and Dumitru-Clementin Cercel. 2020. Approaching smm4h 2020 with ensembles of bert flavours. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 153–157.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of romanian bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4324–4328.

Stefan Daniel Dumitrescu and Andrei-Marius Avram. 2020. Introducing ronec-the romanian named entity corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4436–4443.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, et al. 2020. A report on the vardial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14.

Alexandru-Lucian Georgescu, Horia Cucu, Andi Buzo, and Corneliu Burileanu. 2020. Rsc: A romanian read speech corpus for automatic speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6606–6612.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014a. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014b. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Marius Ionescu, Andrei-Marius Avram, George-Andrei Dima, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020. Upb at fincausal-2020, tasks 1 & 2: Causality analysis in financial documents using pretrained language models. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 55–59.

Akbar Karimi, Leonardo Rossi, Andrea Prati, and Katharina Full. 2020. Adversarial training for aspect-based sentiment analysis with bert. *arXiv preprint arXiv:2001.11316*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. Robert–a romanian bert model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637.

Cristian Onose, Dumitru-Clementin Cercel, and Stefan Trausan-Matu. 2019. Sc-upb at the vardial 2019 evaluation campaign: Moldavian vs. romanian cross-dialect topic identification. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 172–177.

Andrei Paraschiv and Dumitru-Clementin Cercel. 2019. Upb at germeval-2019 task 2: Bert-based offensive language classification of german tweets. In *KONVENS*.

Andrei Paraschiv, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020. Upb at semeval-2020 task 11: Propaganda detection with domain-specific trained bert. *arXiv preprint arXiv:2009.05289*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Cristian Popa and Vlad Ștefănescu. 2020. Applying multilingual and monolingual transformer-based models for dialect identification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 193–201.

Jana Ben Salamah and Aymen Elkhlifi. 2014. Microblogging opinion mining approach for kuwaiti dialect. In *The International Conference on Computing Technology and Information Management (ICCTIM)*, page 388. Society of Digital Information and Wireless Communication.

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. 2014. Sentence level dialect identification for machine translation system selection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 772–778.

Anca Maria Tache, Mihaela Gaman, and Radu Tudor Ionescu. 2021. Clustering word embeddings with self-organizing maps. application on laroseda–a large romanian sentiment data set. *arXiv preprint arXiv:2101.04197*.

Mircea-Adrian Tanase, Dumitru-Clementin Cercel, and Costing-Gabriel Chiru. 2020a. Upb at semeval-2020 task 12: Multilingual offensive language detection on social media by fine-tuning a variety of bert-based models. *arXiv preprint arXiv:2010.13609*.

Mircea-Adrian Tanase, George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020b. Detecting aggressiveness in mexican spanish social media content by fine-tuning transformer-based models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN) 2020*, pages 236–245.

Diana Tudoreanu. 2019. Dteam@ vardial 2019: Ensemble based on skip-gram and triplet loss neural networks for moldavian vs. romanian cross-dialect topic identification. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 202–208.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

George-Alexandru Vlad, George-Eduard Zaharia, Dumitru-Clementin Cercel, Costin-Gabriel Chiru, and Stefan Trausan-Matu. 2020. Upb at semeval-2020 task 8: Joint textual and visual modeling in a multi-task learning architecture for memotion analysis. *arXiv preprint arXiv:2009.02779*.

Nianheng Wu, Eric DeMattos, Kwok Him So, Pin-zhen Chen, and Çağrı Çöltekin. 2019. Language discrimination and transfer learning for similar languages: experiments with feature combinations and adaptation. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 54–63.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Ziqing Yang, Yiming Cui, Zhipeng Chen, Wanxiang Che, Ting Liu, Shijin Wang, and Guoping Hu. 2020. Textbrewer: An open-source knowledge distillation toolkit for natural language processing. *arXiv preprint arXiv:2002.12620*.

George-Eduard Zaharia, Andrei-Marius Avram, Dumitru-Clementin Cercel, and Traian Rebedea. 2020a. Exploring the power of romanian bert for dialect identification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 232–241.

George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020b. Cross-lingual transfer learning for complex word identification. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 384–390. IEEE.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardzic, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, et al. 2019. A report on the third vardial evaluation campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–16.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.