Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 374–384
Marseille, 20-25 June 2022

Distilling the Knowledge of Romanian BERTs Using Multiple Teachers

Andrei-Marius Avram[1], Darius Catrina[*3], Dumitru-Clementin Cercel[2], Mihai Dascalu[2], Traian Rebedea[2], Vasile Păiș[1], Dan Tufiș[1]
Research Institute for Artificial Intelligence, Romanian Academy[1]
University Politehnica of Bucharest, Faculty of Automatic Control and Computers[2]
Duke University[3]
{andrei.avram, vasile, tufis}@racai.ro, {dumitru.cercel, mihai.dascalu, traian.rebedea}@upb.ro

Abstract

Running large-scale pre-trained language models in computationally constrained environments remains a challenging problem yet to be addressed, while transfer learning from these models has become prevalent in Natural Language Processing tasks. Several solutions, including knowledge distillation, network quantization, or network pruning have been previously proposed; however, these approaches focus mostly on the English language, thus widening the gap when considering low-resource languages. In this work, we introduce three light and fast versions of distilled BERT models for the Romanian language: Distil-BERT-base-ro, Distil-RoBERT-base, and DistilMulti-BERT-base-ro. The first two models resulted from the individual distillation of knowledge from two base versions of Romanian BERTs available in literature, while the last one was obtained by distilling their ensemble. To our knowledge, this is the first attempt to create publicly available Romanian distilled BERT models, which were thoroughly evaluated on five tasks: part-of-speech tagging, named entity recognition, sentiment analysis, semantic textual similarity, and dialect identification. Our experimental results argue that the three distilled models offer performance comparable to their teachers, while being twice as fast on a GPU and ∼35% smaller. In addition, we further test the similarity between the predictions of our students versus their teachers by measuring their label and probability loyalty, together with regression loyalty - a new metric introduced in this work.

Keywords: Knowledge Distillation, BERT, Romanian language, Multiple Teachers, Loyalty

1. Introduction

Knowledge transfer from Transformer-based language models (Vaswani et al., 2017) trained on large amounts of data achieves state-of-the-art results on most Natural Language Processing (NLP) tasks (Devlin et al., 2019; Liu et al., 2019; He et al., 2020). However, the best performing models usually have billions (Brown et al., 2020) or even trillions (Fedus et al., 2021) of parameters, making them impractical in certain real-world situations. Moreover, both training and using these language models usually comes at a high environmental cost (Strubell et al., 2019).

Several attempts were made to reduce the size of models by distilling their knowledge (Hinton et al., 2015) accumulated during the pre-training phase (Sanh et al., 2019), after fine-tuning the model on a specific task (Turc et al., 2019), or both pre-training and fine-tuning (Jiao et al., 2020). Other methods consider shrinking the size of the models by either quantizing their weights to integer values (Shen et al., 2020), or pruning parts of the neural network (Brix et al., 2020). In addition, efficient attention mechanisms were developed to overcome the quadratic bottleneck in the sequence length of multi-head attention (Zaheer et al., 2020; Choromanski et al., 2020).

However, the vast majority of these efforts focused on developing English models, and little attention was paid on increasing the efficiency of pre-trained models on other languages, with few singular cases of such compressed models like BERTino (Muffo and Bertino, 2020) for Italian, MBERTA for Arabic (Alyafeai and Ahmad, 2021), or GermDistilBERT[1] for German. As a response to this issue, we focus on Romanian, a language on which BERT has recently attracted a surge of attention from the local community and has shown promising results in various areas like dialect identification (Zaharia et al., 2021; Popa and Ștefănescu, 2020; Zaharia et al., 2020), document classification (Avram et al., 2021) or satire detection in news (Rogoz et al., 2021). Thus, our work introduces three compressed BERT versions for the Romanian language that were obtained through a distillation process:

- **Distil-BERT-base-ro**[2] was obtained by distilling the knowledge of BERT-base-ro (Dumitrescu et al., 2020) using its original training corpus and tokenizer;
- **Distil-RoBERT-base**[3] was created from RoBERT-base (Masala et al., 2020) in similar conditions (i.e., using both original training corpus and tokenizer);

[*]Work done during an interhsip at the Research Institute for Artificial Intelligence, Romanian Academy.

[1]https://huggingface.co/distilbert-base-german-cased

[2]https://huggingface.co/racai/distilbert-base-romanian-cased

[3]https://huggingface.co/racai/distilbert-base-romanian-uncased

- **DistilMulti-BERT-base-ro**[4] considered the distillation of the knowledge from an ensemble consisting of BERT-base-ro and RoBERT-base, while relying on the combined corpus and coupled with the tokenizer of the former model.

Our three compressed models were further evaluated on five Romanian datasets and the results showed that they maintained most of the performance of the original models, while being approximately twice as fast when run on a GPU. In addition, we also measure the label, probability and regression loyalties between each of the three distilled models and their teachers, as quantifying the performance on specific tasks does not show how similar the predictions between a teacher and a student really are. The models, together with the distillation and evaluation scripts were open-sourced to improve the reproducibility of this work[5].

The rest of the paper is structured as follows. The next section presents a series of solutions related to the knowledge distillation of pre-trained language models. The third section outlines our approach of distilling the knowledge of Romanian BERTs, whereas the fourth section presents the evaluation setup and their performance on various Romanian tasks. The fifth section evaluates the prediction loyalty between each distilled version and its teacher, while the sixth section evaluates their inference speed. The final section concludes our work and outlines potential future work.

2. Related Work

Knowledge Distillation (Hinton et al., 2015) is a compression method in which a smaller framework, the student model, is trained to reduce the loss $\mathcal{L}_{KD}$ over the soft probabilities predicted by a larger model (i.e., the teacher):

$$\mathcal{L}_{KD} = \sum_i t_i \cdot log(s_i) \quad (1)$$

where t_i and s_i are the probabilities predicted by the teacher and the student, respectively.

The technique usually uses a temperature parameter T in the softmax function that controls the smoothness of the distribution given by probabilities p_i, known as softmax-temperature:

$$p_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \quad (2)$$

where z_i is the logit corresponding to the probability at index i.

Sanh et al. (2019) introduced the first distilled Transformer-based language model for English, where the authors created a model that was 1.6x smaller, 2.5x faster, and retained 97% of the performance of the original BERT (Devlin et al., 2019) on the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018). Soon after, TinyBERT (Jiao et al., 2020) was introduced and it reduced the size of BERT by 7.5x, improved the inference time by 9.4x, while maintained 96.8% of the performance of the original model on GLUE; TinyBERT performed distillation at both the pre-training and fine-tuning stages. Also, Sun et al. (2020) introduced MobileBERT, a slightly heavier model compared to TinyBERT - 4.3x smaller and 5.5x faster than the original BERT, which managed to retain most of the its knowledge by achieving a GLUE score of 77.7 (0.6 lower than BERT).

One of the first attempts to distill the knowledge of pre-trained language models for languages other than English was BERTino (Muffo and Bertino, 2020) for Italian. The distillation corpus was composed of 1.8 billion tokens and *bert-base-italian-xxl-uncased*[6] was used as the teacher model. The resulting model retained most of the knowledge from the original model, as its performance was below with values ranging from 0.29% to 5.15% on the evaluated tasks.

Additionally, a distilled version of Multilingual BERT (mBERT) - LightMBERT - (Jiao et al., 2021) was developed to reduce the discrepancies between languages by transferring the cross-lingual capabilities of the original model into a smaller one. The authors first initialized the student layers with the bottom layers of the teacher - mBERT. Then, they froze the embedding layer, and performed distillation only on the other layers. This simple process allowed them to maintain the cross-lingual generalization capabilities of the original model.

Nevertheless, the knowledge transferred from a single teacher through distillation may be limited and even biased, resulting in a student model of low quality. As such, Wu et al. (2021) proposed a knowledge distillation approach from several teachers in both the pre-training and fine-tuning stages. Their experimental results showed a significant improvement over single teacher models like DistilBERT or TinyBERT, and even over other models used as teachers.

3. Distillation Process

Figure 1 introduces the overall architecture of the distillation process of all three compressed Romanian BERT variants. The Distil-BERT-base-ro and Distil-RoBERT-base models were obtained by using BERT-base-ro and RoBERT-base as teacher, respectively, together with their teacher's training corpora and tokenizer. DistilMulti-BERT-base-ro used an ensemble of teachers consisting of both Romanian BERTs - BERT-base-ro and RoBERT-base, their combined corpora and the BERT-base-ro tokenizer.

[4] https://huggingface.co/racai/distilbert-multi-base-romanian-cased

[5] https://github.com/racai-ai/Romanian-DistilBERT

[6] https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased

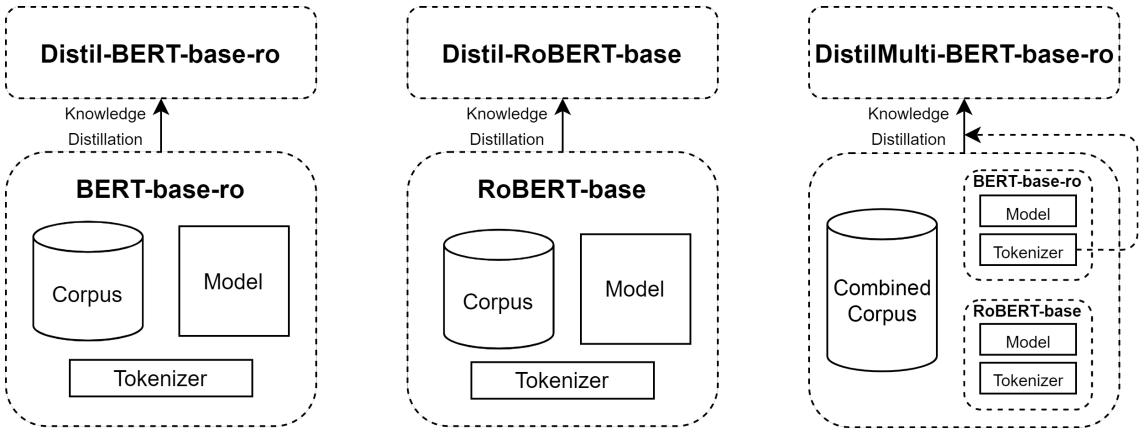

Figure 1: Knowledge distillation process of the three distilled Romanian BERT versions: Distil-BERT-base-ro (left), Distil-RoBERT-base (middle), and DistilMulti-BERT-base-ro (right). DistilMulti-BERT-base-ro uses the BERT-base-ro tokenizer, marked with dotted arrow in the figure.

Corpus	Lines	Words	Size
RoWiki	1.5M	60.5M	0.4 GB
OPUS	55.1M	635M	3.8 GB
OSCAR	33.6M	1725.8M	11 GB
Total	**90.2M**	**2421.3M**	**15.2 GB**

Table 1: BERT-base-ro training corpus.

Corpus	Lines	Words	Size
RoWiki	2M	50M	0.3 GB
RoTex	14M	240M	1.5 GB
OSCAR	87M	1780M	10.8 GB
Total	**103M**	**2070M**	**12.6 GB**

Table 2: RoBERT-base variants training corpus.

3.1. Corpora

Tables 1 and 2 summarize the corpora used for training BERT-base-ro and RoBERT-base, respectively. Both training corpora contain the Romanian Wikipedia and the Open Super-large Crawled Aggregated coRpus (OSCAR) (Suárez et al., 2019) datasets, although their final size is not the same due to preprocessing differences. However, the BERT-base-ro corpus uses the OPUS corpus (Tiedemann, 2012) while RoBERT-base uses the RoTex custom collection[7], resulting in a 2.5 GB difference between the final datasets.

We observed that, in comparison with RoBERT-base training corpus, the BERT-base-ro training corpus was a bit noisier and contained more artifacts; as such, we applied several additional rules to ensure a cleaner version by removing lines in the following conditions: a) lines that contained noise for diacritics (e.g., "*c?nd*" instead of "*când*" - eng., "*when*"); b) lines with uncapitalized versions of the most frequent Romanian named entities (e.g., "*bucurești*" instead of "*București*" - eng.: "*Bucharest*"); and c) lines that were not detected as written in Romanian by `langdetect`[8]. Web page parsing artifacts, such as "*Articolul Anterior*" (eng., "*Previous Article*") at the beginning of the sentence or "Articolul Următor" (eng., "*Next Article*") at the end of the sentence, were also removed. In contrast to the original 15.2 GB of plain text used to train BERT-base-ro, our cleaned version of the corpus contains 14.6 GB of text and approximately 2.3 billions words (i.e., roughly 5.9% smaller).

The combined corpus used to train DistilMulti-BERT-base-ro was obtained by merging and deduplicating the two corpora, resulting in a corpus that contained 25.3 GB of text and 4.1 billion words words.

3.2. Teacher Networks

The teacher models used to perform knowledge distillation during pre-training were the base-cased version of BERT-base-ro (Dumitrescu et al., 2020) for Distil-BERT-base-ro, RoBERT-base (Masala et al., 2020) for Distil-RoBERT-base, and an ensemble of the two for DistilMulti-BERT-base-ro. Although, to the best of our knowledge, two other Romanian BERT models exist, namely RoBERT-small and RoBERT-large that were introduced in the same work as RoBERT-base (Masala et al., 2020), we only chose these variants because their parameters configuration best fit the compression process used to obtain DistilBERT.

3.3. Student Networks

All the distilled versions introduced in this work follow the architecture of DistilBERT, having 6 Transformer encoder layers (i.e., half the layers of BERT-base-ro), 12 attention heads and a hidden dimension of 768. All follow-up analyses consider as reference also mBERT (Devlin et al., 2019), the most accessible multi-lingual BERT-based model that supports Roma-

[7] `https://github.com/aleris/ReadME-RoTex-Corpus-Builder`

[8] `https://github.com/Mimino666/langdetect`

nian; XLM-RoBERTa (XLM-R) (Conneau et al., 2020) was not considered due to its size which is one order of magnitude larger than the other models. Table 3 outlines the size, the number of parameters, and several layers statistics (i.e., the number of layers and heads, and the hidden size) for mBERT, RoBERT, BERT-base-ro, and the distilled models. Both Distil-BERT-base-ro and DistilMulti-BERT-base-ro have a size of 312 MB and contain 81 millions parameters, reducing the size of BERT-base-ro by ∼35%. Distil-RoBERT-base is slightly smaller, with 72 millions parameters and a size of 282 MB, compressing RoBERT-base by the same amount[9].

Following Muffo and Bertino (2020), the total loss function $\mathcal{L}$ used to pre-train Distil-BERT-base-ro and Distil-RoBERT-base was composed of three parts $\mathcal{L}_{KD}$, $\mathcal{L}_{MLM}$ and $\mathcal{L}_{COS}$, weighted by different coefficients that sum to one:

$$\mathcal{L} = \lambda_{KD}\mathcal{L}_{KD} + \lambda_{MLM}\mathcal{L}_{MLM} + \lambda_{COS}\mathcal{L}_{COS} \quad (3)$$

where $\mathcal{L}_{KD}$ is the knowledge distillation loss (Hinton et al., 2015), $\mathcal{L}_{MLM}$ is the masked language modeling (MLM) loss (Devlin et al., 2019), $\mathcal{L}_{COS}$ is the cosine similarity embedding loss used to align the hidden states of the teacher and student models (Sanh et al., 2019), and λ_{KD}, λ_{MLM} and λ_{COS} are the weights of each loss. A higher weight for the knowledge distillation loss $\lambda_{KD} = 0.625$ is considered because the training corpora for the student network represents a large portion (∼94%) of the pre-training corpora for the teacher; as such, more of the internal knowledge acquired by the teachers would be transferred to the students in the process, at the detriment of not learning much by itself based on the MLM loss. λ_{MLM} and λ_{COS} are set to 0.25 and 0.125, respectively.

The knowledge distillation loss $\mathcal{L}_{KD}$ and the cosine similarity loss $\mathcal{L}_{COS}$ were split for training DistilMulti-BERT-base-ro into two equally weighted parts corresponding to the losses of each model in the ensemble :

$$\mathcal{L}_{KD} = \frac{\mathcal{L}^1_{KD} + \mathcal{L}^2_{KD}}{2} \quad (4)$$

$$\mathcal{L}_{COS} = \frac{\mathcal{L}^1_{COS} + \mathcal{L}^2_{COS}}{2} \quad (5)$$

where $\mathcal{L}^1_{KD}$, $\mathcal{L}^1_{COS}$ are the distillation and cosine similarity losses corresponding to BERT-base-ro, and $\mathcal{L}^2_{KD}$, $\mathcal{L}^2_{COS}$ are the distillation and cosine similarity losses corresponding to RoBERT-base.

3.4. Training Settings

An important practical aspect of knowledge distillation is the initialization of the student parameters. Following the setup of DistilBERT, we initialize the students layers with the teacher layers by taking out the first layer of two[10] (i.e. first layer, third layer, fifth layer etc.). The distilled models were trained for 3 epochs with a batch size of 256 and a learning rate of 5e-4. We also applied a weight decay of 1e-4 and a warm-up for the first 5% of the entire training process, together with gradient clipping (Pascanu et al., 2013) for gradient norms surpassing the value of 5. The temperature T from Equation 2 was set to 2. The training process took approximately 30 days on two GeForce GTX 1080 Ti for each model variant.

4. Task Evaluation

Five Romanian tasks were considered to create a strong evaluation setup for our distilled models:

- **Part-of-Speech (POS) Tagging**: Label a sequence of tokens with Universal Part-of-Speech (UPOS) and eXtended Part-of-Speech (XPOS).
- **Named Entity Recognition (NER)**: Tag a sequence of tokens with Inside–Outside–Beginning (IOB) labels (Ramshaw and Marcus, 1999).
- **Sentiment Analysis (SA)**: Predict whether a review expresses a positive or negative sentiment (SAPN), together with its rating (SAR).
- **Dialect Identification (DI)**: Identify the Romanian/Moldavian dialects in news articles.
- **Semantic Textual Similarity (STS)**: Given a pair of sentences, measure how semantically similar they are.

The fine-tuning of the models for each task was performed by using the AdamW optimizer (Loshchilov and Hutter, 2017) and a scheduler that linearly decreased the learning rate to 0 at the end of the training. The number of epochs, the batch size, the learning rate, and the warm-up steps used for each individual task are depicted in Table 4. In line with the recommendations of the authors of the STS evaluation scripts[11], we did not employ any warm-up steps, a higher batch size of 256 was considered, and early stopping was used for training instead of a maximum number of epochs. Each experiment was run 5 times and we report the average scores in order to mitigate the variation in performance due to the random initialization of the weights.

4.1. Part-of-Speech Tagging

Task Description. The original splits of the Romanian Reference Trees (RRT) corpus (Barbu Mititelu et al., 2016) from Universal Dependencies (UD) v2.7 were used to train and evaluate the distilled models on UPOS and XPOS tagging. The evaluation metric employed to measure the performance on both subtasks was the macro-averaged F1-score.

[9]The difference in the number of parameters between Distil-BERT-base-ro and RoBERT-base is generated by the difference in vocabulary size.

[10]We initialized DistilMulti-BERT-base-ro with the parameters of BERT-base-ro.

[11]https://github.com/dumitrescustefan/RO-STS/tree/master/baseline-models

Model	Layers	Hidden	Heads	Vocab	Size	Params
mBERT	12	768	12	120K	681 MB	177M
BERT-base-ro	12	768	12	50K	477 MB	124M
RoBERT-small	12	256	8	38K	74 MB	19M
RoBERT-base	12	768	12	38K	441 MB	114M
RoBERT-large	24	1024	16	38K	1.3 GB	341M
Distil-BERT-base-ro	6	768	12	50K	312 MB	81M
Distil-RoBERT-base	6	768	12	38K	282 MB	72M
DistilMulti-BERT-base-ro	6	768	12	50K	312 MB	81M

Table 3: Model size comparison of mBERT, Romanian BERTs, and our distilled versions.

Task	Epochs	Batch	Warm	L-Rate
UPOS	10	16	1000	1e-4
XPOS	10	16	1000	4e-5
NER	15	16	500	5e-5
SAPN	10	16	1000	3e-5
SAR	10	16	1000	5e-5
DI	5	8	1500	5e-5
STS	-	256	0	2e-5

Table 4: Hyperparamters used to fine-tune the models on each evaluation task.

Model	UPOS	XPOS
mBERT	97.87	96.16
RoBERT-small	97.43	96.05
RoBERT-base	98.02	97.15
RoBERT-large	**98.12**	**97.81**
BERT-base-ro	98.00	96.46
Distil-BERT-base-ro	97.97	97.08
Distil-RoBERT-base	97.12	95.79
DistilMulti-BERT-base-ro	98.07	96.83

Table 5: UPOS and XPOS evaluation results on RRT.

Methodology. The output embeddings of the model E_i were projected into a tensor of dimension equal with the number of classes of UPOSes or XPOSes. The transformation function was a feed-forward layer with weights W, bias b, and the $LeakyReLU$ activation function that produced the logits y_i corresponding to each embedding (i.e., $y_i = LeakyReLU(W^T E_i + b)$); a dropout of 0.1 was also applied for regularization. The logits were further transformed using the $softmax$ function to obtain the output distribution. The cross-entropy loss between the target labels and the predicted probabilities was employed as the value to be minimized.

Results Model performance on UPOS and XPOS prediction are outlined in Table 5. The highest F1-score of 98.07% on UPOS evaluation was achieved by DistilMulti-BERT-base-ro, outperforming both its teachers and being the second model in the overall leaderboard behind RoBERT-large. The best XPOS performance was obtained by Distil-BERT-base-ro with 97.08%, surpassing BERT-base-ro with 96.46%, but falling behind its teacher and the large BERT variant.

An in-depth analysis of Romanian UPOS and XPOS evaluation was performed in (Păiş et al., 2021), where several basic language processing kits (BLARK) were tested on RRT. Our distilled models obtained superior results on UPOS and comparable results with the best models on XPOS.

4.2. Named Entity Recognition

Task Description. NER evaluation was performed on Romanian Named Entity Corpus (RONEC) (Dumitrescu and Avram, 2020). Models were evaluated according to Segura-Bedmar et al. (2013) and the macro-averaged F1-scores are reported for the exact matches of the IOB labels.

Methodology. The approach used to fine-tune the models is the same as the one used in POS tagging, the only difference being the dimension of the output tensor that was adjusted to match the number of classes found in RONEC.

Results The results on this task are presented in Table 6. The Distil-BERT-base-ro obtained a strict F1-score of 79.42% which is almost identical to the strict F1-score obtained by the distilled ensemble - 79.43%, both models outperforming Distil-RoBERT-base by approximately 0.3% on this metric. Our compressed models lagged behind the base models by more than 3% on the strict F1 metric (i.e., RoBERT-large and BERT-base-ro) and by more than 2.5% on the exact F1-score (i.e., BERT-base-ro), but they managed to achieve a performance close to the teachers on the other two metrics.

4.3. Sentiment Analysis

Task Description SA was performed on the Large Romanian Sentiment Data Set (LaRoSeDa) (Tache et al., 2021), a dataset that contains 15,000 reviews written in Romanian, of which 7,500 are positive and 7,500 negative. We fine-tuned the models to predict both positive and negative sentiments, as well as the rating of each review. The models were evaluated by measuring the accuracy and the F1-macro score for the two prediction subtasks.

Model	Type	Partial	Strict	Exact
mBERT	84.52	86.27	80.60	84.13
RoBERT-small	83.11	84.59	78.61	82.06
RoBERT-base	85.92	87.21	82.05	85.14
RoBERT-large	**86.45**	87.19	**82.61**	85.09
BERT-base-ro	86.21	**87.84**	82.54	**85.88**
Distil-BERT-base-ro	83.83	85.73	79.42	83.35
Distil-RoBERT-base	85.80	87.39	79.15	83.11
DistilMulti-BERT-base-ro	85.48	87.66	79.43	83.22

Table 6: NER evaluation results on RONEC.

Methodology The model was given as input two sentences - the title of the review and the content of the review - separated by the token `[SEP]`. We projected the embedding C of the token `[CLS]` into a scalar y, representing the sentiment of the review, by using a linear layer with weights W and bias b, to which the $LeakyReLU$ activation function is applied (i.e., $y = \sigma(W^T C + b)$). Then, the $sigmoid$ function is applied to the scalar y to obtain the output probability. For the rating prediction, the input was also composed of the title and the content of the review separated by the `[SEP]` token. However, the embedding corresponding to the `[CLS]` token was projected using a feed-forward neural network into four dimensions representing the logits of possible ratings. Also, the $softmax$ function is employed instead of the $sigmoid$ function to transform the output of the neural network. The models were trained to reduce the binary cross-entropy for SAPN and the cross-entropy losses for SAR, respectively.

Results Table 7 depicts the results of the SA task. DistilRo-BERT achieved the best performance on the positive versus negative sentiment analysis out of all other distilled models, with 98.20% accuracy and a 98.12% F1-score. For rating prediction, DistilRo-BERT obtained an accuracy score of 90.14% and a F1-score of 80.51% surpassing all the other evaluated models. The difference in scores between the sentiment and the rating prediction for each model might be due to the noisy labeling of the ratings, as stated by the authors of the dataset (Tache et al., 2021).

4.4. Dialect Identification

Task Description The MOldavian and ROmanian Dialectal COrpus (MOROCO) (Butnaru and Ionescu, 2019) is a dataset that contains 33,564 samples of text that were collected from news, annotated with their dialect and with one of six topics that each sample belongs to. We evaluated the proposed models in terms of accuracy and F1-score (macro averaged) on the identification of Romanian versus Moldavian dialects on MOROCO samples.

Methodology The same approach as in the previously described positive versus negative sentiment analysis being a binary classification task was considered for this task. The only slight difference is that the MOROCO samples do not have a title; as such, no separation of the title and the content was necessary.

Results DI results are outlined in Table 8. As it can be observed, Distil-BERT-base-ro and DistilMulti-BERT-base-ro achieve better performance than their teachers on this task (i.e., accuracy/F1-scores of 96.36%/96.31% and 96.26%/96.16%), outperforming even the large variant of RoBERT.

4.5. Semantic Textual Similarity

Task Description Romanian Semantic Textual Similarity (RoSTS)[12] is a dataset that was obtained by translating the Semantic Textual Similarity (STS) dataset[13] and used to compare the capacity of the employed BERT models in capturing the semantic similarity between Romanian sentences. The evaluation metrics are the Pearson and the Spearman coefficients.

Methodology The fine-tuning of the language models was performed by giving as input the two sentences separated by the `[SEP]` token, followed by the projection of the resulted `[CLS]` token embedding into a scalar y using a linear layer with weights W and bias b, to which the $LeakyRELU$ function is applied (i.e., $y = \sigma(W^T C + b)$). Then, the $sigmoid$ function is applied to the output logit. To match with the y interval, the original similarities were normalized from the $[0, 5]$ to $[0, 1]$. The models were trained using the mean squared error loss.

Results Task results are outlined in Table 9. DistilMulti-BERT-base-ro achieved the highest average Pearson/Spearman scores (80.66%/80.27%) out of all distilled models, slightly outperforming Distil-BERT-base-ro by approximately 0.1%/0.25% and Distil-RoBERT-base by approximately 0.85%/0.25%. In comparison with the teacher models, both Distil-BERT-base-ro and DistilMulti-BERT-base-ro achieved better results than BERT-base-ro, but lagged behind RoBERT-base by approximately 0.5% on both metrics and RoBERT-large by approximately 2%/1.5%.

[12] https://github.com/dumitrescustefan/RO-STS

[13] https://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark

Model	SAPN		SAR	
	Acc	F1	Acc	F1
mBERT	97.43	97.28	88.12	78.98
RoBERT-small	97.37	97.22	87.55	77.81
RoBERT-base	**98.30**	**98.20**	89.69	79.40
RoBERT-large	98.25	98.16	89.91	79.82
BERT-base-ro	98.07	97.94	88.45	79.61
Distil-BERT-base-ro	98.20	98.12	**90.14**	**80.51**
Distil-RoBERT-base	98.01	97.61	88.33	79.58
DistilMulti-BERT-base-ro	98.11	97.74	89.43	79.77

Table 7: SA evaluation results for positive and negative and rating classification on LaRoSeDa.

Model	Acc	F1
mBERT	95.12	95.06
RoBERT-small	95.48	95.41
RoBERT-base	96.10	96.07
RoBERT-large	96.20	96.17
BERT-base-ro	95.64	95.58
Distil-BERT-base-ro	**96.36**	**96.31**
Distil-RoBERT-base	96.13	96.11
DistilMulti-BERT-base-ro	96.26	96.18

Table 8: DI evaluation results on MOROCO.

Model	Pears	Spear
mBERT	76.64	76.41
RoBERT-small	78.24	77.84
RoBERT-base	81.18	80.69
RoBERT-large	**82.51**	**81.83**
BERT-base-ro	80.30	79.94
Distil-BERT-base-ro	80.57	80.02
Distil-RoBERT-base	79.80	78.82
DistilMulti-BERT-base-ro	80.66	80.27

Table 9: STS evaluation results on RoSTS.

5. Loyalty

Xu et al. (2021) introduced two new metrics - loyalty and robustness - for measuring the similarities between a student and a teacher model because just comparing the evaluation metric for a specific task does not reflect how alike the student and the teacher model behave. For loyalty, the authors measure similarities between the labels (L-L) and the probabilities (P-L) predicted by the models fine-tuned on Multi-Genre Natural Language Inference dataset (Williams et al., 2018). However, we choose to evaluate the loyalties of our distilled models on the test set of LaRoSeDa as there is no natural language inference dataset for Romanian. In addition, we introduce a new loyalty evaluation metric for measuring the similarity for regression - regression loyalty (R-L) -, using the test set of RoSTS as target. Robustness was measured by Jin et al. (2020) using the after-attack accuracy (AA); however, it is not used in this work due to the lack of high-quality word embeddings that are retrofitted for synonymy[14]. The loyalty metrics for MultiDistil-BERT-base-ro were computed using the average of the metrics computed with each individual teacher.

Model	L-L	P-L	R-L
Distil-BERT-base-ro	87.80	**74.76**	94.05
Distil-RoBERT-base	84.63	71.95	92.24
DistilMulti-BERT-base-ro	**89.75**	73.23	**94.64**

Table 10: Label, probability and regression loyalty results between our distilled models and their teachers.

5.1. Label Loyalty

Label Loyalty (L-L) directly measures the similarity between the labels predicted by two models as the accuracy between the teacher labels acting as the ground truth versus the student labels:

$$LL = Accuracy(label_t, label_s) \tag{6}$$

where $label_t$ and $label_s$ are the labels predicted by the teacher and the student, respectively.
The highest L-L score was obtained by DistilMulti-BERT-base-ro with 89.75%, as outlined in Table 10, showing that the distilled ensemble has superior label alignment with its teachers than each individual distilled model.

5.2. Probability Loyalty

Computing the similarities between the output probabilities of a teacher and student after compression is also important in industrial applications that focus on confidence and meaningfulness (Guo et al., 2017). P-L evaluates the distance between the teacher output probabilities P_t and the student output probabilities P_s as:

$$PL = 1 - \sqrt{D_{JS}(P_t||P_s)} \tag{7}$$

where D_{JS} is the Jensen–Shannon divergence, defined using the Kullback–Leibler divergence over probabilities X $D_{KL}(P||Q) = \sum_{x \in X} P(x) log(\frac{P(x)}{Q(x)})$ as:

[14]It must be noted that a version of synonymy aware word embeddings was introduced in (Dumitrescu et al., 2018), but they were not open-sourced for public usage.

$$D_{JS}(P_t||P_s) = \frac{D_{KL}(P_t||P_s) + D_{KL}(P_s||P_t)}{2} \quad (8)$$

The results of P-L evaluation are depicted in the second column of Table 10. The highest score was obtained by Distil-BERT-base-ro with 74.76%, outperforming the distilled ensemble model with approximately 1.5%.

5.3. Regression Loyalty

The previous two metrics do not assess performance in the case of regression problems. Thus, this work introduces regression loyalty as the Pearson Correlation Coefficient (Breese et al., 1998) between the output of the teacher $pred_t$ and the output of the student $pred_s$:

$$RL = Pearson(pred_t, pred_s) \quad (9)$$

The results for R-L are outlined in the last column of Table 10. DistilMulti-BERT-base-ro achieved the highest regression loyalty with its two teachers (a 94.64% R-L score), followed by Distil-BERT-base-ro (94.05%) and Distil-RoBERT-base (92.24%).

6. Inference Speed

The inference time of the three distilled versions in comparison with the other Romanian or multilingual models was evaluated on random sequences with lengths ranging from 16 tokens to 512 tokens, using both a CPU - Intel i7-7700K - and a GPU - GeForce GTX 1080 Ti. The models were grouped by their size into four categories to make the visualization more appealing, namely: Multilingual (i.e., mBERT), Small (i.e., RoBERT-small), Base (i.e., BERT-base-ro and RoBERT-base), Large (i.e., RoBERT-large), and Distilled (i.e., Distill-BERT-base-ro, Distill-RoBERT-base, and DistillMulti-BERT-base-ro). The inference times are depicted in Figure 2.

The distilled models obtained a significant improvement on the GPU, being almost twice as fast as the base and small models, and at least three times faster than the RoBERT-large. It should be noted that the small variant obtains a similar speed on the GPU for larger sequence lengths due to more parallelization happening at the level of the input tokens; RoBERT-small benefits from its reduced size and overcomes its disadvantage in parallelization by having 12 layers instead of 6. The inference times on the CPU of the distilled models is also better than those obtained by the base and large models, but worse than those of RoBERT-small because it has less parameters.

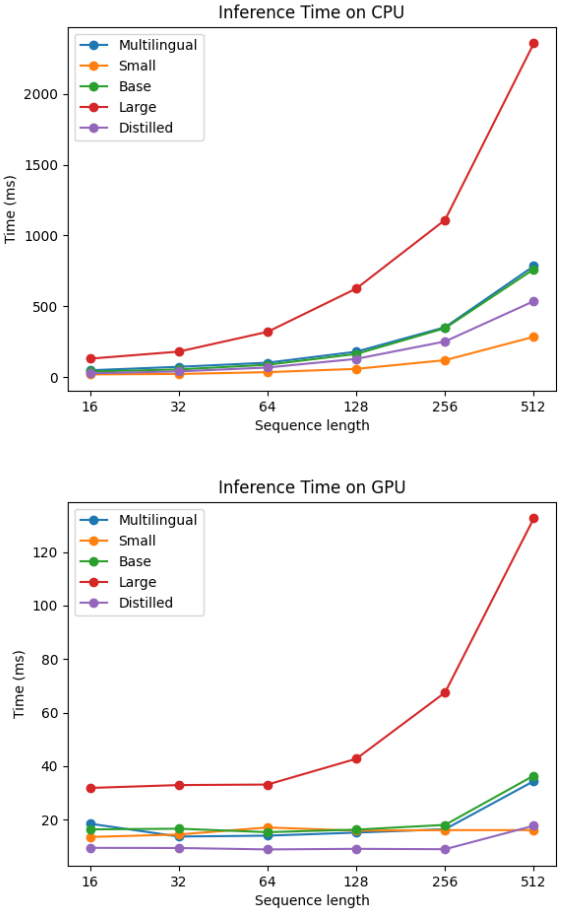

Figure 2: Inference time on CPU (up) and on GPU (down) of the evaluated models, grouped by their dimensions: Multilingual (mBERT), Small (RoBERT-small), Base (BERT-base-ro, RoBERT-base), Large (RoBERT-large) and Distilled (Distill-BERT-base-ro, Distill-RoBERT-base, DistillMulti-BERT-base-ro).

7. Conclusions

Creating models for low resource languages is an important area of research, aiming to empower the usage of NLP in various contexts. This work introduces three distilled models from Romanian BERT models: Distil-BERT-base-ro, Distil-RoBERT-base, and DistilMulti-BERT-base-ro, that are 35% smaller than their teacher models (BERT-base-ro and RoBERT-base). Each model was evaluated on five Romanian NLP tasks: part-of-speech tagging, named entity recognition, sentiment analysis, dialect identification, and semantic textual similarity. Our experimental results showed that the distilled models maintain most of the prediction performance of their original models, while containing only half of their layers. Moreover, the distilled models even outperformed their teachers and other evaluated models on several tasks. We further tested the inference speed of the three distilled models on various sequence lengths, and our results outlined a reduction of almost 50% on a single GPU in comparison to their teacher models.

Future work considers the creation of a distilled version using RoBERT-large as teacher, as well as further reducing the size of the current models by pruning and/or quantizing their weights. In addition, we intend to fine-tune the distilled models and annotate the Romanian sub-corpus from the CURLICAT multilingual

corpus covering domains relevant for CEF Digital Service Infra-structures (DSIs). Besides standard mark-up (lemmatization, POS tagging, chunking, dependency parsing) annotation will include recognition and labeling of the occurrences of terms recorded in the Interactive Terminology for Europe (IATE) [15], s described in the Curated Multilingual Language Resources for CEF.AT (CURLICAT) [16] project. Finally, the results presented in this paper are going to be included on the five evaluation tasks in LiRo (Dumitrescu et al., 2021), a benchmark for Romanian language NLP tasks.

Acknowledgements

This research was supported by the EC grant INEA/CEF/ICT/A2018/28592472 for the Action No: 2019-EU-IA-0034 entitled "Curated Multilingual Language Resources for CEF.AT" (CURLICAT) and by the Romanian Ministry of European Investments and Projects through the Competitiveness Operational Program (POC) project "HOLOTRAIN" (grant no. 29/221, SMIS code: 129077).

8. Bibliographical References

Alyafeai, Z. and Ahmad, I. (2021). Arabic compact language modelling for resource limited devices. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 53–59.

Avram, A.-M., Păiș, V., and Tufis, D. I. (2021). Pyeurovoc: A tool for multilingual legal document classification with eurovoc descriptors. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 92–101.

Barbu Mititelu, V., Ion, R., Simionescu, R., Irimia, E., and Perez, C.-A. (2016). The romanian treebank annotated according to universal dependencies. In *Proceedings of the tenth international conference on natural language processing (hrtal2016)*.

Breese, J. S., Heckerman, D., and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 43–52.

Brix, C., Bahar, P., and Ney, H. (2020). Successfully applying the stabilized lottery ticket hypothesis to the transformer architecture. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3909–3915.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Butnaru, A. and Ionescu, R. T. (2019). Moroco: The moldavian and romanian dialectal corpus. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 688–698.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al. (2020). Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dumitrescu, S. D. and Avram, A.-M. (2020). Introducing ronec-the romanian named entity corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4436–4443.

Dumitrescu, S. D., Avram, A. M., Morogan, L., and Toma, S.-A. (2018). Rowordnet–a python api for the romanian wordnet. In *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–6. IEEE.

Dumitrescu, S., Avram, A.-M., and Pyysalo, S. (2020). The birth of romanian bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4324–4328.

Dumitrescu, S. D., Rebeja, P., Lorincz, B., Gaman, M., Avram, A., Ilie, M., Pruteanu, A., Stan, A., Rosia, L., Iacobescu, C., Morogan, L., Dima, G., Marchidan, G., Rebedea, T., Chitez, M., Yogatama, D., Ruder, S., Ionescu, R. T., Pascanu, R., and Patraucean, V. (2021). Liro: Benchmark and leaderboard for romanian language tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Fedus, W., Zoph, B., and Shazeer, N. (2021). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

He, P., Liu, X., Gao, J., and Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L.,

[15] https://iate.europa.eu/home

[16] https://curlicat.eu/

Wang, F., and Liu, Q. (2020). Tinybert: Distilling bert for natural language understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4163–4174.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. (2021). Lightmbert: A simple yet effective method for multilingual bert distillation. *arXiv preprint arXiv:2103.06418*.

Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. (2020). Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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

Muffo, M. and Bertino, E. (2020). Bertino: An italian distilbert model. In *CLiC-it*.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. PMLR.

Popa, C. and Ștefănescu, V. (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.

Păiş, V., Ion, R., Avram, A.-M., Mitrofan, M., and Tufiș, D. (2021). In-depth evaluation of Romanian natural language processing pipelines. *Romanian Journal of Information Science and Technology (ROMJIST)*, 24(4):384–401.

Ramshaw, L. A. and Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Rogoz, A.-C., Gaman, M., and Ionescu, R. T. (2021). Saroco: Detecting satire in a novel romanian corpus of news articles. *arXiv preprint arXiv:2105.06456*.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Segura-Bedmar, I., Martínez, P., and Herrero-Zazo, M. (2013). Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350.

Shen, S., Dong, Z., Ye, J., Ma, L., Yao, Z., Gholami, A., Mahoney, M. W., and Keutzer, K. (2020). Q-bert: Hessian based ultra low precision quantization of bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8815–8821.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650.

Suárez, P. J. O., Sagot, B., and Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., and Zhou, D. (2020). Mobilebert: a compact task-agnostic bert for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170.

Tache, A., Mihaela, G., and Ionescu, R. T. (2021). Clustering word embeddings with self-organizing maps. application on laroseda–a large romanian sentiment data set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 949–956.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218.

Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Wu, C., Wu, F., and Huang, Y. (2021). One

teacher is enough? pre-trained language model distillation from multiple teachers. *arXiv preprint arXiv:2106.01023*.

Xu, C., Zhou, W., Ge, T., Xu, K., McAuley, J., and Wei, F. (2021). Beyond preserved accuracy: Evaluating loyalty and robustness of bert compression. *arXiv preprint arXiv:2109.03228*.

Zaharia, G.-E., Avram, A.-M., Cercel, D.-C., and Rebedea, T. (2020). 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.

Zaharia, G.-E., Avram, A.-M., Cercel, D.-C., and Rebedea, T. (2021). Dialect identification through adversarial learning and knowledge distillation on romanian bert. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 113–119.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al. (2020). Big bird: Transformers for longer sequences. In *NeurIPS*.