Cross-Lingual and Multi-Task Learning with Knowledge Distillation for Emotion Recognition in Low-Resource Romanian

Vlad-Cristian Matei
Department of Computer Science
National University of Science and Technology POLITEHNICA
Bucharest, Romania
vlad_cristian.matei@stud.acs.upb.ro

Abstract

This study is expected to advance emotion recognition in the Romanian language, addressing the challenges caused by limited datasets and methodologies. In this respect, cross-lingual domain adaptation was done by leveraging datasets in English with the purpose of enabling transfer learning to compensate for the resource scarcity.

Further improvements were obtained using a multi-task learning framework: jointly performing emotion recognition, sentiment analysis, and news categorization tasks. Self-knowledge distillation and teacher annealing further optimized the shared learning process. The results indicate good progress in emotion recognition for Romanian texts, setting a solid foundation for future advances in low-resource languages.

1 Introduction

Emotions color every experience we live, representing a central aspect of relationships between people. Research in this area has intensified, involving various fields such as: psychology, humanities and social sciences. Emotion analysis is a task that has been intensively researched in the Natural Language Processing (NLP) community for years [7]. In her book [1], Dr. Rosalind Picard described automatic emotion recognition as: *"Giving Computers Emotional Skills"*. Such systems can be incredibly powerful: facilitators of enormous progress (commerce), but also perpetrators of great harm (manipulating voters). However, [14] provides a complex analysis regarding the ethics behind emotion analysis and provides several reasons why automatic emotion analysis remains relevant and very important.

The growing popularity of emotions analysis has sparked an explosion of datasets, methodologies, and academic papers in recent years. However, most successful models have been developed and tested in resource-rich environments—providing extensive training time and meticulously annotated texts. When it comes to the Romanian language, however, the landscape is entirely different. The availability of datasets dedicated to emotions analysis is severely limited, the amount of available data being quite small [5]. Furthermore, there have been few attempts to explore a diverse range of approaches specifically tailored to the complexity of the Romanian language.

The main goal of this study is to improve the fields of emotion, sentiment, and news categorization analysis, in the context of the Romanian language, which currently still suffers from the lack of

[1] Picard, Rosalind W. 2000. Affective Computing. MIT Press. https://doi.org/10.7551/mitpress/1140.001.0001

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

resources and diverse methodologies. To achieve this, three main tasks are considered: (i) *emotion recognition*, (ii) *sentiment analysis*, (iii) *news categorization*. Each of these tasks will be trained separately, utilizing a technique of cross-lingual domain adaptation. The underlying idea behind this solution is that existing resources in the English language can improve models trained on Romanian through transfer learning. This will involve utilizing high-resource English datasets from different domains that target the same tasks. This combination will not only help in achieving the primary objectives but will also help clarify our understanding about the relationships between Romanian and English.

In the second phase, I will explore, within a multi-task learning framework, to what extent the tasks are influenced by one another. This will involve conducting an ablation study, with the expectation that at least one combination will yield improvements in the final results. To ensure that the information is effectively harmonized in this more complex multi-tasking environment, I will also use a technique of self-knowledge distillation along with a teacher annealing method. All results will be evaluated through the analysis of accuracy, precision, recall and F_1 score.

In short, this work brings the following contributions: (i) three models trained on each of the three tasks (emotion recognition, sentiment analysis, news categorization); (ii) those models being augmented on English language datasets via cross-lingual domain adaptation, presenting some intermediate results as well; and (iii) combining the initial finetuned models within a multi-tasking framework along with the self-knowledge distillation and teacher annealing methods, and analyzing the results.

2 Related Work

This section summarizes related work, highlighting significant models, datasets, and methods that motivated this work. I also describe how the methods contrast with these previous models to give a unique approach that makes sense both practically and theoretically.

2.1 Model

The introduction of BERT in the field of NLP [12] has been described as revolutionary. Pretrained on massive multi-language corpora, it tops leaderboards across a variety of tasks such as text classification and also shows remarkable cross-lingual transfer capabilities. "The birth of Romanian BERT" [8] introduced a new variant of BERT specifically adapted for Romanian, called BERT-base-ro. This model was pretrained on approximately 15 GB of Romanian text, significantly improving the ability of BERT-based models to capture the linguistic little nuances of the Romanian language. Thus, in this work, models will mainly rely on BERT-base-ro, as the embeddings represent English words, while being adapted well for Romanian. This provides an excellent way to explore the cross-lingual transfer of information.

2.2 Datasets

REDv2 [2] [5] is a large available resource for **emotion analysis** in Romanian language. Its second version includes 5,449 manually verified tweets, each marked with one of these seven emotions: *Anger (17.8%)*, *Fear (12%)*, *Joy (15.6%)*, *Sadness (20%)*, *Surprise (10.6%)*, *Trust (11.5%)*, and *Neutral (25.6%)*. The original split - 75% train, 10% validation, and 15% test will be adopted for use in this research. The SOTA on this dataset is found in [5], results obtained by Ro-BERT, published with a F_1 of 0.668 and by XLM-Roberta with a 0.619 F_1. As this will be the primary dataset for this research, I will try to exceed such figures.

The secondary data sets utilized fall into two categories: Romanian tasks and their corresponding English tasks.

Thus, the dataset that will enrich the primary task through its contribution is the **Emotion dataset** [16]. It is made of 20.000 English twitter messages categorized into six basic emotions: *Anger (13.5%)*, *Fear (12.1%)*, *Joy (33.5%)*, *Love (8.2%)*, *Sadness (29.2%)*, and *Surprise (3.6%)*. The original split - 80% for training, 10% for validation and 10% for testing - will proceed in this research to be used.

[2] `https://github.com/Alegzandra/RED-Romanian-Emotions-Dataset/tree/main`

I will use the pair LaRoSeDa [18] [3] - SST2 [17] [4] for the **sentence analysis task**. **LaRoSeDa** is a dataset containing 15,000 reviews collected from one of the largest e-commerce platforms in Romania. Each piece of text can be clearly tagged as *Positive (50%)* or *Negative (50%)*, enabling binary sentiment differentiation. For training, I have selected 11.000 examples, while for testing, I used only 3.000, and for validation, 1.000. **SST-2** is a corpus that contains sentences from film reviews, marked as *Positive (55.8%)* or *Negative (44.2%)*. Approximately split, 96% was used for training, 1.3% for validation, and 2.6% for testing. It can be observed that multiple domains are used, not only in terms of language but also when thinking about the source of the datasets.

For the **news categorization task**, I will use the MoRoCo [2] [5] - Ag News [20] pair. [6]. **MoRoCo** is a compilation of news articles crawled from the Romanian and Moldovian websites. These sentences are labeled into six topics: *Culture (6.8%)*, *Finance (25.4%)*, *Politics (27.2%)*, *Science (8.7%)*, *Sports (18%)* and *Tech (13.9%)*. For testing, I will use a split of 64.7% train, 17.6% test and 17.6% for validation. **Ag News** is a large news corpus, containing more than 1 million examples. It is build by the news search engine ComeToMyHead; however, this research will utilize only a subset of 120,000 examples for training and 7,600 for testing. This dataset has four top level labels: *World (25%)*, *Sports (25%)*, *Business (25%)* and *Sci/Tech (25%)* .

2.3 Methods

Many studies have examined how tasks can be enhanced by exposing models to different distributions, tasks, and domains, taking advantage of transfer learning techniques. In a comprehensive survey [15], transfer learning is categorized into three main types: inductive, transductive, and unsupervised. In this research, the method I use to enhance the emotion classification task in one domain (Romanian) by using resources from another domain (English) for the same task falls under **transductive learning**, with the mention that, while limited, there are labeled data available in Romanian.

Moreover, as a specific subcategory of transductive learning, **cross-lingual learning** [1] comes into play, particularly since both languages have some labeled data it can be called even poly-lingual learning as defined in [1]. Thus, the tasks specific to the Romanian language will be enhanced by leveraging the corresponding tasks in English, benefiting from the transfer of information. In cases where the domains in each language also differ, the problem can be viewed as one of **cross-domain learning** as well.

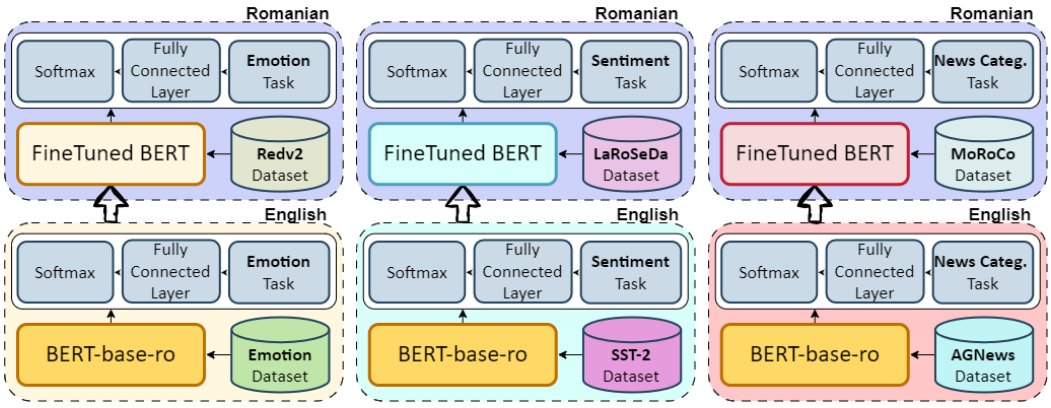

Figure 1 Multi-Stage Cross-Lingual FineTuning

In [19], the authors demonstrate that better results can be achieved simply by applying **multi-stage fine-tuning** to a pre-trained model, a domain adaptation technique. Their method involves a gradual transition from a source domain to the target domain, progressively increasing the proportion of target domain examples until it becomes the sole focus in the final training stage. Since this method is straightforward to implement, I will use it in this research to evaluate the effectiveness of the other techniques, with the mention that I will apply simple training, without successive recursive steps.

[3] https://huggingface.co/datasets/universityofbucharest/laroseda

[4] https://huggingface.co/datasets/stanfordnlp/sst2

[5] https://huggingface.co/datasets/universityofbucharest/moroco

[6] https://huggingface.co/datasets/fancyzhx/ag_news

As a cross-lingual classification method, the **Adversarial Deep Averaging Networks (ADAN)** [4] introduces a unified framework for transferring information from a high-resource language (English) to a low-resource language (Chinese). ADAN leverages a shared feature extractor that enhances the sentiment classifier while simultaneously harming the language discriminator, whose task is to identify whether the input text is from the source or target language. If the discriminator fails to differentiate between languages, the extracted features become effectively language-invariant. Inspired by this approach, the aim of this work is to implement a BERT-based architecture to address both classification and language transfer tasks. I will translate the English dataset into Romanian using Google Translate and by using backpropagation, the model should be improved as following the adversarial training paradigm outlined in ADAN. Briefly, it will be multi-task learning: starting with a model fine-tuned on an English dataset, followed by multi-task learning on both the Romanian dataset and a language classification task between sentences in the original English domain dataset and their corresponding translations in Romanian.

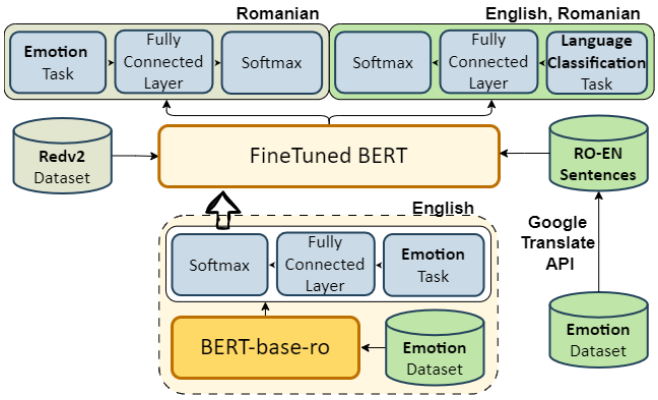

Figure 2 Multi-Task Cross-Lingual Same-Domain Arhitecture

Multi-Task Learning (MTL) [15] is a paradigm in which multiple related tasks can be trained simultaneously in a unified framework and are eventually expected to improve the performance of the primary task. By using a shared architecture, the weights of the neurons are updated together, acting as a regularizer. Furthermore, information is intuitively interrelated, so the more tasks the model is taught, the more competent it will be, just like the student learning from several teachers [3]. Such multifaceted exposure to a rich diversity of perspectives and insights will provide the model with in-depth understanding and better performances on all tasks.

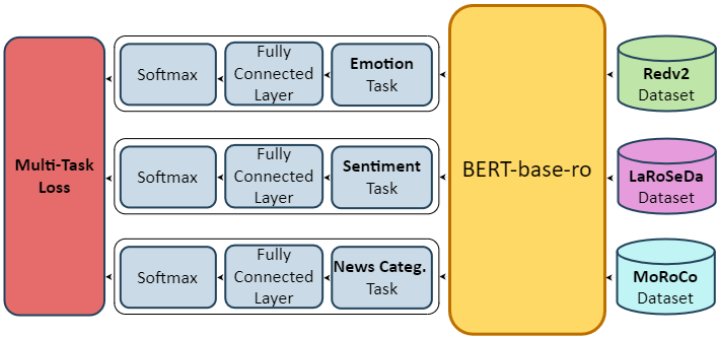

Figure 3 Multi-Task Learning Arhitecture

Since it is often not straightforward for a model to learn various tasks simultaneously, the authors in [6] present "Born-Again Multi-Task Networks," a distilled model based on the same architecture as the original, benefiting from a multi-tasking framework. The technique of **Knowledge Distillation** [9] is a method that transfers knowledge from an expert model, called the teacher, to an initial model, called the student, for the specific task at hand. This relies on the student being able to replicate the responses of the teacher and take cues from it. The advantage comes because the teacher's responses, with their distribution across the final classes, is more signal than a one-hot label. [9] refers to this information as "dark knowledge." In this article, I will use all of the technique presented in [6], which also includes the use of teacher annealing [11]. By maintaining an architecture of similar size, it is

presumed that a comparable learning capacity is preserved, while also placing greater emphasis on how this information is learned. I will refer to the final model obtained as a **Multi-Task Knowledge Distilled (MTKD)** model.

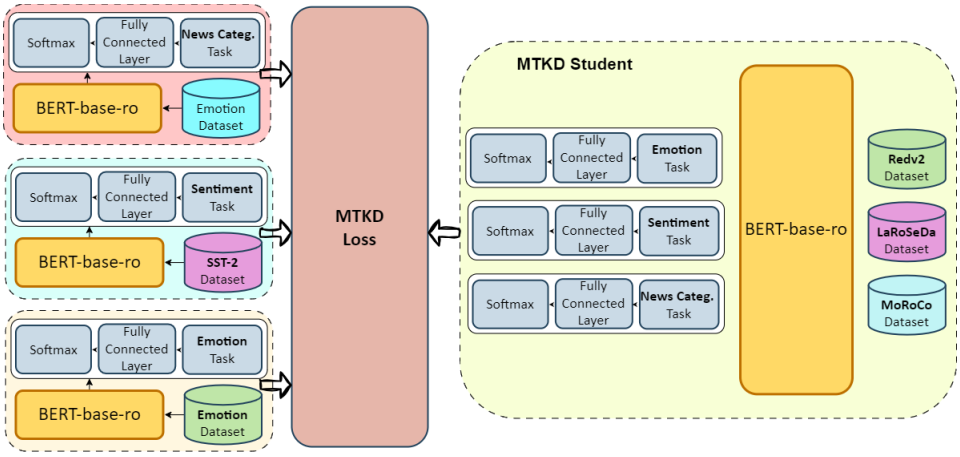

Figure 4 Multi-Task Knowledge Distillation Arhitecture

Teacher Annealing (TA) [11] is a method from knowledge distillation which mitigates the increasing gap between the teacher and the student models. With the growth in the capacity of the teacher model, its influence could be harmful to the student's performance [13]. TA combats this by slowly ramping down the influence of the teacher, while increasing the weight on true labels during training. This is done by adjusting a parameter α from 0 to 1, that balances the distillation loss and the supervised loss. This approach helps the student model be more independent and perform better.

3 Experimental Setup

The experimental setup which includes the methodologies, loss functions, and hyperparameters to evaluate the models.

3.1 Performance Metrics

Various evaluation metrics were used to evaluate model performance objectively:

- **Accuracy:**
$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$
- **Precision:**
$$\text{Precision} = \frac{TP}{TP + FP}$$
- **Recall:**
$$\text{Recall} = \frac{TP}{TP + FN}$$
- **F$_\beta$ score:**
$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}}$$
For this evaluation, we used $\beta = 1$.

3.2 Loss Functions Overview

This section will go over the loss functions utilized for the experiments, providing a brief overview of each along with their purpose for the model training and performance.

3.2.1 Binary Cross-Entropy Loss

The function `BCEWithLogitsLoss` was used for multi-label classification tasks. It combines Binary Cross-Entropy (BCE) with a $sigmoid$ activation, converting logits into probabilities. The formula is:

$$L_{BCE} = -\frac{1}{N \cdot K} \sum_{i=1}^{N} \sum_{k=1}^{K} [y_{i,k} \cdot \log(p_{i,k}) + (1 - y_{i,k}) \cdot \log(1 - p_{i,k})]$$

where $y_{i,k}$ is the one-hot encoded ground truth for the k-th class, and $p_{i,k}$ is the model's predicted probability. This loss was applied in fine-tuning tasks and multi-task settings across datasets in the same domain.

3.2.2 Multi-Task Learning Loss

Let $D_\tau = \{(x_{\tau i}, y_{\tau i})\}$ represent the training dataset for task τ. The multi-task loss, $\mathcal{L}_{MTL}$, is computed as follows [10]:

$$\mathcal{L}_{MTL}(\theta) = \sum_{\tau=1}^{3} \sum_{(x_{\tau i}, y_{\tau i}) \in D_\tau} \mathcal{L}_{BCE}(y_{\tau i}, f_\tau(x_{\tau i}; \theta))$$

Here, $f_\tau(x_{\tau i}; \theta)$ represents the model's output for task τ given input $x_{\tau i}$ and parameters θ.

3.2.3 Multi-Task Knowledge Distillation Loss

Knowledge distillation is a process wherein knowledge flows from a teacher model to a student model. The temperature parameter T plays the central role in this process through smoothing of the output probabilities of the teacher model, where high T encourages exploration by distributing probability mass across classes, while low T sharpens the predictions toward dominant classes. The probability of class k, $p_{i,k}$ is computed using the following temperature-scaled softmax [9]:

$$p_{i,k} = \frac{\exp(z_{i,k}/T)}{\sum_j \exp(z_{i,j}/T)},$$

where $z_{i,k}$ is the teacher model's logit for class k.

The distillation framework combines two losses: 1. Binary Cross-Entropy (BCE), which measures the discrepancy between the student model's predictions and the ground truth. 2. Kullback-Leibler (KL) divergence, which aligns the student's soft predictions with the teacher's softened output. The KL divergence-based distillation loss is:

$$L_{\text{KL-KD}} = T^2 \cdot \frac{1}{N} \sum_{i=1}^{N} \text{KL}\big(p_t(x_i, T) \| p_s(x_i, T)\big),$$

where N is the dataset size, and $p_t(x_i, T)$, $p_s(x_i, T)$ are the soft probability distributions from the teacher and student, respectively.

The combined loss function is:

$$L_{\text{KD}} = \alpha L_{\text{BCE}} + (1 - \alpha) L_{\text{KL-KD}},$$

where α balances the contributions of BCE and KL divergence.

For multi-task knowledge distillation (MTKD), the framework extends to multiple tasks. The MTKD loss for a task τ is computed over its dataset $\mathcal{D}_\tau$:

$$L_{\text{KL-MTKD}}(\theta) = \sum_{\tau=1}^{3} T^2 \cdot \frac{1}{N_\tau} \sum_{(x_{\tau i}, y_{\tau i}) \in \mathcal{D}_\tau} \text{KL}\big(f_\tau(x_{\tau i}; \theta_\tau) \| f_\tau(x_{\tau i}; \theta)\big),$$

where: $f_\tau(x_{\tau i}; \theta_\tau)$ and $f_\tau(x_{\tau i}; \theta)$ are the teacher and student outputs for task τ, N_τ is the number of samples in task τ, $\mathcal{D}_\tau$ is the dataset for task τ, θ_τ and θ are the parameters of the teacher and student models.

The final MTKD loss is:

$$L_{\text{MTKD}} = \alpha L_{\text{BCE}} + (1 - \alpha) L_{\text{KL-MTKD}},$$

where α determines the trade-off between aligning with the ground truth (BCE) and distilling teacher knowledge (KL-MTKD). This formulation will enable us to do multi-task learning effectively by transferring the knowledge from task-specific teacher models to a unified multi-task student model [11, 10].

3.2.4 Teacher-Annealing in Multi-Task Knowledge Distillation

The MTKD-TA loss function is largely similar to the MTKD due to the presence of the teacher annealing mechanism. In this case, parameter α makes a balance between BCE loss and KL divergence loss. Varying α from 0 to 1 modifies how much the model depends upon the teacher's knowledge versus the model's own predictions.

The values of α used in the experiments ranged from 0 to 1, from 0.2 to 0.7, and 0.8 in order to see the effects that it has on the model's performance.

3.3 Hyperparameter Configuration

This section outlines the hyperparameters used for the experiments. This work primarily focus on emotion detection, so further details can be found in the attached notebooks.

For all experiments, the batch size is kept at 16, and weight decay is set as 0.01. In fine-tuning BERT for emotion detection, the model was trained for 7 epochs at a learning rate of 2×10^{-5}. After pre-training on the Emotions Dataset, the model was further fine-tuned using RedV2 with a learning rate of 1.7×10^{-5} for 4 epochs: the multi-stage fine-tuning phase. The learning rate was set to 3×10^{-5} for 4 epochs in the multi-task cross-lingual same-domain experiment. For multi-task learning across multiple domains, the model used a learning rate of 2×10^{-5} for 4 epochs.

In order to account for the fact that the LaRoSeDa and MoRoCo datasets contain many more examples than the others, approximately 10 times more, the loss relative to these two tasks is scaled down by a factor of 10, whereas the loss for emotion detection remains unchanged.

For the MTKD experiments, I consider a learning rate of 2×10^{-5} for 3 epochs. Also, various temperatures values used are 0.5, 1, 1.5, 2, 3, and 4. The α parameter has been varied at values 0.5, 0.75, and 1 to balance the contributions of each task to the final loss.

During the experimentation with MTKD-TA, the α hyperparameter was tested in the range of 0.1–0.7, while the temperature values were considered as 1, 1.5, and 2. Again, the learning rate was kept at 2×10^{-5} for 3 epochs.

Finally, the configuration adopted in the experiments of MTKD-TA is kept for the ablation study, focused on the best performing configuration, namely, the MTKD-TA setup.

4 Results

In this section, the results of the experiments are discussed, together with an interpretation of the findings. However, most of the time, the focus is set on emotion recognition, despite the fact that the scores from the other tasks are provided in the tables. This attempt is done to logically reason, based on the results and to get an in-depth view from the effects of the usage of the methods.

4.1 Evaluation of Cross-Lingual Enhancements

The results obtained will be now presented, together with an analysis of how the model is able to improve its predictions, considering that the main dataset, RedV2, has only 5,449 examples. Moreover, it is taken into consideration that BERT-base-ro was pre-trained on a total of 15GB of text. Further on, potential outcomes will be explored regarding the fact that BERT-base-ro originates from pre-training on a multilingual multi-BERT model, including knowledge of both Romanian and English.

On the main task, the results reflect some improvements, albeit small and mostly in low-resource settings. In cases where the dataset is larger and already performs well—like Sentiment Analysis

Table 1: Performance Results for Cross-Lingual Approaches

Model	Emotion Recognition				Sentiment Analysis				News Categorization			
	Ac	F_1	Pr	Re	Ac	F_1	Pr	Re	Ac	F_1	Pr	Re
base models												
[SotA] Ro-BERT	54.10	66.80	-	-	-	-	-	-	-	-	-	-
BERT-base-ro	56.84	67.33	69.40	65.48	95.23	95.28	95.27	95.30	84.85	86.92	87.10	86.79
cross-lingual settings												
Multi-Stage Cross-Lingual	57.82	**68.05**	73.44	63.53	95.23	95.28	95.27	95.30	85.51	86.90	87.32	86.59
Multi-Task Cross-Lingual	57.94	**68.07**	71.55	65.15	95.33	**95.36**	95.39	95.33	85.28	86.69	87.30	86.54

(close to 95%) or News Categorization (86%)—no notable improvement was achieved by adding pre-training on a dataset in another language, even with the benefits of the multi-BERT model. While on lower baselines, like 67.33%, that leave room for more incrementation, there was a very decent 0.7% increase. Despite falling within the same domain of emotion recognition and emotion analysis in social media (Twitter), the optimization methods, including multi-stage cross-lingual fine-tuning and multi-task cross-lingual same-domain fine-tuning, did not work effectively.

Although initial expectations were improvements through cross-lingual multi-task learning, those did not occur, likely for a variety of reasons. One possible explanation lies in that the language discrimination task English vs. Romanian was just too simple, meaning the model overfitted in just a few iterations, leaving small room for meaningful improvements. Also, there was no use of any contrastive loss or advanced technique for rearranging the logits in the dimensional space. During the experiments, different settings were tried, such as using a loss sign reversal to introduce confusion instead of discrimination, so as to align the correlations between the English and Romanian datasets. This setting also failed to provide significant improvements.

4.2 Evaluation of Multi-Task and Knowledge Distillation Enhancements

Table 2 presents the results for the primary task achieved under the optimal configuration, leveraging a multi-tasking framework enhanced with knowledge distillation and teacher annealing.

Table 2: Performance Results for Multi-Task and Knowledge Distillation approaches

Model	Emotion Recognition				Sentiment Analysis				News Categorization			
	Ac	F_1	Pr	Re	Ac	F_1	Pr	Re	Ac	F_1	Pr	Re
base models												
[SotA] Ro-BERT	**54.10**	**66.80**	-	-	-	-	-	-	-	-	-	-
BERT-base-ro	56.84	67.33	69.40	65.48	95.23	95.28	95.27	95.30	84.85	86.92	87.10	86.79
Best of Cross-Lingual	**57.94**	68.07	**71.55**	65.15	95.33	95.36	95.39	95.33	85.28	86.89	87.30	86.54
multi-task settings												
Multi-Task Learning	57.57	**68.53**	70.73	**66.55**	95.10	95.18	95.16	95.20	83.96	85.81	86.31	85.33
Multi-Task Knowledge-Distillation	56.11	**69.11**	70.59	**67.74**	**95.53**	**95.56**	**95.59**	**95.53**	81.60	**87.33**	85.03	**90.40**
MTKD and Teacher Annealing	51.58	**70.19**	66.98	**74.00**	95.26	*95.40*	95.37	95.43	78.20	86.37	81.73	**92.60**

While the model should theoretically be more capable, as learning diverse information from multiple teacher models helps, the multi-tasking environment is inherently complex. First, there are many differences among the datasets that make it hard for learning to take place. Second, the contribution of each task might be different; some tasks can be favored in training processes while others become overshadowed.

Table 3 highlights experiments that reduce the contribution of auxiliary tasks by dividing their loss by 5, 10, or 15. Indeed, the best setting corresponds to dividing the loss by 10, which is consistent with the fact that auxiliary datasets are about ten times larger than the primary RedV2 dataset. However, this aggressive reduction caused secondary tasks to perform worse since less emphasis was placed on their optimization.

Results in **Table 2** confirm the expected trade-off: whereas the auxiliary tasks have stabilized or slightly regressed, the main task of emotion recognition showed measurable improvements. Compared to the BERT-base-ro baseline, performance increased by 1.2% with multi-task learning, 1.8% with multi-task learning and knowledge distillation, and 2.9% with adding teacher annealing process.

Successive improvement in the results points out the contribution of each technique: multitask learning provides a positive effect, knowledge distillation brings the advantage of using "dark knowledge,"

and teacher annealing, where during training the student model progressively relies more on its predictions, further improves the results.

Further supporting the robustness of these learning techniques are the auxiliary tasks that showed improvements: sentiment analysis had a final gain of about 0.3%, and news categorization had an improvement of 0.4%. These results, one after another, show how effective and adaptive the applied methodologies have been.

Table 3: Impact of Multi-Task Loss Calibration on Performance Across Multiple Tasks

Model	Loss / Div	Emotion Recognition				Sentiment Analysis				News Categorization			
		Ac	F_1	Pr	Re	Ac	F_1	Pr	Re	Ac	F_1	Pr	Re
multi-task loss calibration													
Multi-Task Learning	5	54.64	**66.79**	65.82	67.96	95.46	95.50	95.46	95.53	83.79	84.69	85.27	84.13
Multi-Task Learning	10	57.57	**68.53**	**70.73**	66.55	95.10	95.18	95.16	95.20	83.96	85.81	86.31	85.33
Multi-Task Learning	15	55.86	**67.74**	69.04	66.55	95.33	95.40	95.33	95.46	84.01	86.12	87.10	85.21

4.2.1 Hyperparameter Exploration in Multi-Task Learning and Knowledge Distillation Frameworks

Next, I will present a few observations that can be deduced from the hyperparameter exploration process.

Table 4: Impact of Hyperparameter Tuning on Emotion Recognition

Model	Alpha	Temperature	Emotion Recognition			
			Ac	F_1	Pr	Re
multi-task knowledge distillation hypertunning						
Proposed model	**0.1-0.7**	**1.5**	**51.58**	**70.19**	**66.98**	**74.00**
MTKD	0.5	0.5	54.52	67.58	**74.72**	61.81
	0.5	1.0	56.11	**69.11**	70.59	67.74
	0.5	1.5	50.36	**68.88**	64.57	73.89
	0.5	2.0	48.16	68.63	62.06	**76.91**
	0.5	3.0	48.41	68.21	62.09	**75.83**
	0.5	4.0	47.43	67.78	61.58	**75.51**
MTKD	0.75	1.0	56.23	68.39	72.02	65.15
	0.75	1.5	55.99	68.96	70.03	67.96
	0.75	2.0	55.99	69.03	69.29	68.82
MTKD	1.0	1.0	57.21	67.83	70.54	65.48
	1.0	2.0	56.23	66.87	69.72	64.29
MTKD-TA	0.1-0.7	1.0	56.11	69.07	74.53	64.50
	0.1-0.7	2.0	48.41	68.76	63.77	75.08

It can be inferred from **Table 4** that a correlation can be established between the adjustment of the α and T hyperparameters with respect to their effect on the performance metrics. In the first section, where α is set to 0.5, it is evident that with the rise in temperature, the recall score improves while the precision score decreases. However, the highest accuracy is achieved where the best F_1 score occurs, representing a trade-off between too high and too low temperatures, specifically around a value of 1.0.

In the second section, for α equals 0.75, the same behavior is observed. Comparing these sections, it is obvious that only at a T around 2.0 will a similar result to the best one, from the first section, be obtained. In other words, the more relying on hard predictions and ground truth, distillation will contribute, but in such a way that the classified information is smoother, allowing for greater exploration and less confidence in sharp predictions. Therefore, knowledge distillation has a less decisive influence, acting more as advice rather than a determining factor.

The third section shows that a model which only relies on ground truth, and therefore does not benefit from distillation, does not give better results, even when the temperature is varied.

Lastly, the final section confirms the benefit of gradually shifting from reliance on teacher confidence to the student's independent decision-making. By restricting the α range to a maximum of 0.7, we ensure that the student always considers, but to a lesser extent, the "advice" received from the teacher. It is also evident here that the best combination occurs with the proposed final model, where a temperature of 1.5 proves to be the winning solution.

4.3 Ablation Study

Next, an ablation study will be presented, allowing the behavior of the model to be observed when certain tasks are excluded from the multi-tasking framework.

Table 5: Ablation Study on the Multi-Task Framework

Model	Emotion Recognition			
	Ac	F_1	Pr	Re
w/o - without task				
Proposed model	51.58	70.19	66.98	74.00
w/o sentiment analysis	54.52	69.37	67.87	70.98
w/o news categorization	55.50	68.37	69.72	67.20

As can be seen in **Table 5**, the results demonstrate the importance of each task in improving the model's ability to correctly recognize emotions within a text. Surprisingly, the task of news categorization seems to have a greater and more positive influence on the model than the sentiment analysis task, which appears to be more closely related to the main task. Thus, the initial hypothesis is confirmed, with all tasks contributing effectively to the primary objective.

4.4 Qualitative Analysis

Considering the final results, since this is a multi-label classification problem, the results are still far from qualitative. The final F_1 score is only around 70%, which is still pretty low and means that the predictions are aligned with the measured performance.

Table 6: Qualitative Analysis

Text	Label	Prediction (%)
Premierul <|PERSON|> mizează pe "convingerea românilor de a respecta regulile de izolare la domiciliu	Trust	Suprise (59), Neutral (54), Fear (53) , Anger (54), Trust (52)
Ce am și eu fobie socială aiurea	Fear	Suprise (61), Neutral (58), Anger (51)
EXCLUSIV! Este dezastru național! Nu mai poate fi declanșată starea de urgență în România! <|PERSON|> și PNL aruncă țara-n haos! <|URL|>… prin	Fear	Anger (58), Surprise (57), Fear (54), Trust (54), Neutral (52)
Omul binedispus <|URL|>…	Joy	Trust (58), Anger (55)

As it can be observed, in the first example, the model gives several labels, and words like "convingerea" (conviction) keep suggesting *Trust*, whereas "izolare" (isolation) would suggest *Anger*. Therefore, I would not say that the understanding of the model is somehow far from being correct.

In the second example, the lack of punctuation probably misled the model in predicting *Surprise*, but the confusion between *Fear* and *Anger* does not seem very serious.

However, in the last two examples, the model still seems to be confused and is unable to make the right decisions.

5 Conclusions

This paper concludes by successfully developing a model that can recognize various emotions within Romanian text-a low-resource language for training in the main task.

Starting from an initial F_1 score of just 67.33%, indicating room for improvement, several optimization techniques have been tried. The first approach involved using English-language datasets from identical or slightly different domains and leveraging them within a cross-lingual framework. The best technique in this attempt improved the F_1 score to 68.06%, which indicates that this method, although promising, needs to be adjusted, as the progress, though present, is not substantial.

Then, a multi-task learning framework was applied, which brought more significant improvements. By solving the sentiment analysis and news categorization tasks concurrently, the result was enhanced to 68.53%. Moreover, after applying knowledge distillation and teacher annealing, the scores further increased to 69.11% and 70.19%, respectively. Given its success, this approach was analyzed in greater detail by testing some hyperparameter adjustments.

The ablation study showed that all auxiliary tasks are contributing positively towards the gain in performance. Moreover, the qualitative analysis also showed that while the model improves, it still fails in many cases, which provides room for further improvements.

I personally think that the study fulfilled its goals since it provided relevant conclusions and also pointed out the need for future directions that must include new techniques to improve these results even further.

References

[1] Nuria Bel, Cornelis Koster, and Marta Villegas. Cross-lingual text categorization. volume 18, 04 2004.

[2] Andrei M. Butnaru and Radu Tudor Ionescu. Moroco: The moldavian and romanian dialectal corpus, 2019.

[3] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

[4] Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570, 2018.

[5] Alexandra Ciobotaru, Mihai Vlad Constantinescu, Liviu P. Dinu, and Stefan Dumitrescu. RED v2: Enhancing RED dataset for multi-label emotion detection. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1392–1399, Marseille, France, June 2022. European Language Resources Association.

[6] Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, and Quoc V. Le. BAM! born-again multi-task networks for natural language understanding. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5931–5937, Florence, Italy, July 2019. Association for Computational Linguistics.

[7] Flor Miriam Plaza del Arco, Alba Curry, Amanda Cercas Curry, and Dirk Hovy. Emotion analysis in nlp: Trends, gaps and roadmap for future directions, 2024.

[8] Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. The birth of Romanian BERT. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online, November 2020. Association for Computational Linguistics.

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

[10] Mahshid Hosseini and Cornelia Caragea. Distilling knowledge for empathy detection. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3713–3724, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[11] Aref Jafari, Mehdi Rezagholizadeh, Pranav Sharma, and Ali Ghodsi. Annealing knowledge distillation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2493–2504, 2021.

[12] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[13] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI Conference on Artificial Intelligence*, 2019.

[14] Saif M. Mohammad. Ethics sheet for automatic emotion recognition and sentiment analysis, 2022.

[15] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359, 2010.

[16] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[17] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[18] Anca Tache, Gaman Mihaela, and Radu Tudor Ionescu. Clustering word embeddings with self-organizing maps. application on LaRoSeDa - a large Romanian sentiment data set. pages 949–956, April 2021.

[19] Haoran Xu, Seth Ebner, Mahsa Yarmohammadi, Aaron Steven White, Benjamin Van Durme, and Kenton Murray. Gradual fine-tuning for low-resource domain adaptation, 2021.

[20] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification, 2016.