Cross-Lingual and Multi-Task Learning with Knowledge Distillation for Emotion Classification 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

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, [12] 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 totally 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 primary aim of this research is to advance the fields of emotion, sentiment, and news categorization analysis, particularly in the context of the Romanian language, which currently faces a lack of resources and diverse methodologies. To achieve this, three basic tasks are considered: (i) *emotion recognition*, (ii) *sentiment analysis*, (iii) *news categorization*. Each of these tasks will be trained individually, utilizing a technique of cross-lingual domain adaptation. The rationale behind this approach is that existing resources in the English language can enhance models trained on Romanian by transfer learning. This will involve leveraging rich English datasets from various domains that address the same tasks. This combination will not only aid in achieving the primary objectives but will also enhance our understanding of the relationships between Romanian and English.

In the second phase, I will investigate, within a multi-task learning framework, how the tasks influence 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

[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).

harmonized in this more complex multi-tasking environment, I will also employ 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 f1 score.

To summarize, the contributions of this work are: (i) training three models, each tailored to one of the three tasks (emotion recognition, sentiment analysis, news categorization); (ii) enhancing these models with English language datasets through cross-lingual domain adaptation, along with the presentation of intermediate results; and (iii) once the final models have been obtained, combining them within a multi-tasking framework along with the self-knowledge distillation and teacher annealing methods, and analyzing the results.

2 Related Work

2.1 Model

BERT [11] has significantly transformed the field of NLP. It was pretrained on large-scale corpora in multiple languages, allowing it to achieve state-of-the-art results across a variety of tasks, such as text classification and it has demonstrated impressive 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 nuances of the Romanian language. Thus, in this study, the models will be primarily based on BERT-base-ro, whose embeddings are capable of understanding English, while being effectively fine-tuned for Romanian. This creates an optimal environment for exploring cross-lingual transfer of information.

2.2 Datasets

REDv2 [2] [5] is a publicly available resource for **emotion analysis**. In its second version, it contains 5,449 manually verified tweets, each categorized into one of 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 used in this research. The SOTA on this dataset is reported in [5], with an F1 score of 0.668 for Ro-BERT and 0.619 for XLM-Roberta. As the primary dataset for this study, I will aim to surpass this score.

The secondary datasets utilized can be divided into two categories: tasks in Romanian and their corresponding tasks in English.

Thus, the dataset that will enrich the primary task through its contribution is the **Emotion dataset** [14]. This dataset consists 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 be employed in this research.

For the **sentiment analysis task**, I will utilize the LaRoSeDa [16] [3] - SST2 [15] [4] pair. **LaRoSeDa** is a dataset that includes 15,000 reviews collected from one of the largest e-commerce platforms in Romania. Each text can be classified as *Positive (50%)* or *Negative (50%)*, enabling binary sentiment classification. For training, I have retained 11.000 examples, while for testing, I used 3.000, and for validation, 1.000. **SST-2** is a corpus composed of sentences extracted from movie reviews, labeled as *Positive (55.8%)* or *Negative (44.2%)*. Regarding the split, approximately 96% was designated for training, 1.3% for validation, and 2.6% for testing. It can be observed that there are various domains, not only in terms of language but also regarding the source of the datasets.

For the **news categorization task**, I will use the MoRoCo [2] [5] - Ag News [18] pair. [6]. **MoRoCo** consists of a collection of sentences gathered from news articles published on websites in Romania and Moldova. These sentences are classified into six topics: *Culture (6.8%)*, *Finance (25.4%)*, *Politics*

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

[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

(27.2%), *Science (8.7%)*, *Sports (18%)* and *Tech (13.9%)*. For testing, I will utilize a split of 64.7% train, 17.6% test and 17.6% for validation. **Ag News** is a vast collection of news articles, comprising over 1 million examples. It was created 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. In this dataset, there are four main 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 [13], 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.

In [17], 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.

As a cross-lingual sentiment 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, I aim 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.

Multi-Task Learning (MTL) [13] is a methodology that enables the simultaneous training of multiple related tasks within a cohesive framework, ultimately aimed at enhancing the performance of the primary task. By employing a shared architecture, the weights of the neurons are updated concurrently, functioning as a regularization technique. Moreover, the information is intuitively interconnected; thus, by learning multiple tasks, the model becomes more capable, much like a student who enhances their abilities by learning from several teachers [3]. This multifaceted exposure allows the model to absorb diverse perspectives and insights, leading to a more robust understanding and improved performance across tasks.

Since it can be challenging for a model to learn multiple tasks simultaneously, the authors in [6] present "Born-Again Multi-Task Networks," a distilled model built upon the same architecture as the original, benefiting from a multi-tasking framework. The technique of **Knowledge Distillation** [9] involves transferring information from a specialized model, referred to as the teacher, to an initiating model, known as the student, in the context of that specific task. This approach relies on the student's ability to mimic the teacher's responses and be influenced by them. The advantage arises from the fact that the teacher's responses, with their distribution across the final classes, provide more signal than a one-hot label. [9] describes this information as "dark knowledge." In this article, I will employ the entire technique presented in [6], which also involves the use of teacher annealing [10]. 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.

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] 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.

[11] 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.

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

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

[14] 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.

[15] 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.

[16] 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.

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

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