Computer Science & Engineering Department

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

Matei Vlad Cristian
AI Master II

University Politehnica of Bucharest

Why this study for Romanian language ?

Unsaturated Research

Insufficient researchers to advance new discoveries for the Romanian language. Opportunity and potential exist for new experiments.

Low-Resource Language

As the availability of datasets dedicated to emotions analysis is severely limited, the exploration of diverse range of approaches is needed.

What are the objectives ?

01

Train **BERT-base-ro** on three tasks on Romanian language: *(i) emotion recognition, (ii) sentiment analysis, (iii) news categorization.* Enrich them with a cross-lingual domain adaptation method by incorporating English datasets.

02

Within a **multi-task learning** framework, investigate how the tasks influence one another. To ensure that the information is effectively harmonized, a technique of **self-knowledge distillation** with **teacher annealing** will be used.

Related work

The Architecture

2.1 Model

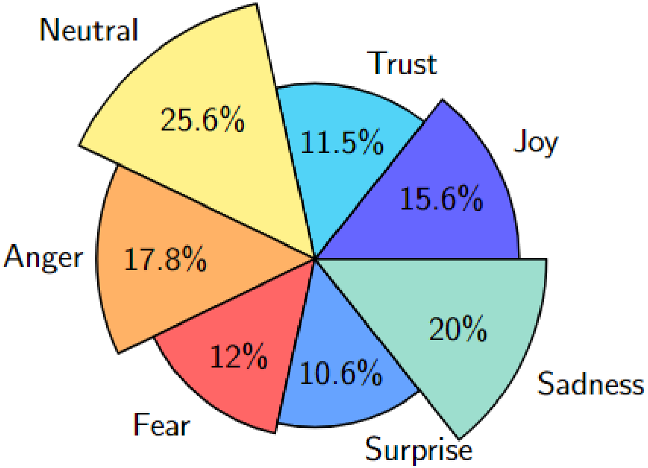
BERT [11] has significantly transformed the field of NLP. It was pretrained on large-scale corpora in multiple languages, allowing it to achieve state-of-the-art results across a variety of tasks, such as text classification and it has demonstrated impressive cross-lingual transfer capabilities. "The 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.

The Datasets

01 Datasets Establishment

Redv2

Emotions Task

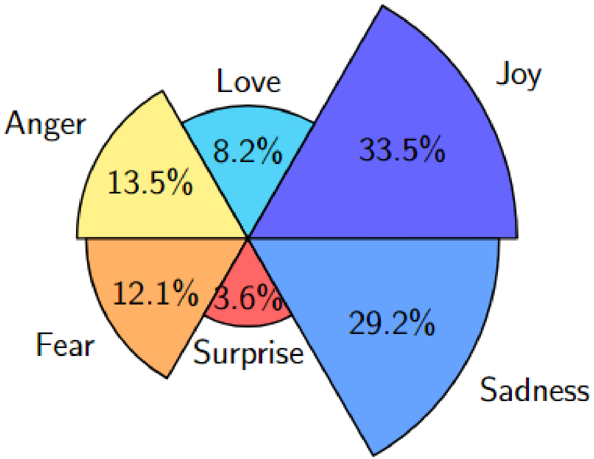


- It contains 5,449 manually verified tweets
- The original split - 75% train, 10% validation, and 15% test will be used in this research.
- SOTA: 0.668 Ro-BERT F1.

01 Datasets Establishment

Emotion Dataset

Emotions Task

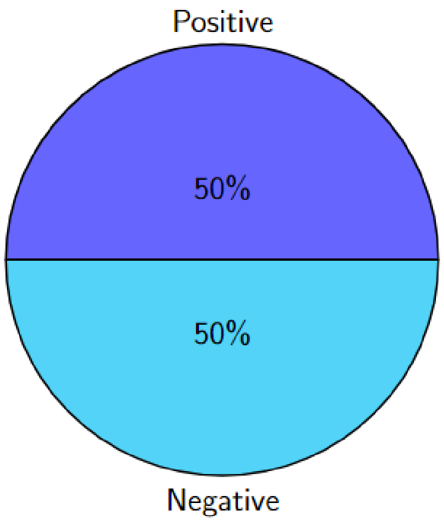


- It contains 20K English tweets
- The original split - 80% train, 10% validation, and 10% test will be used in this research.

02 Datasets Establishment

LaRoSeDa Dataset

Sentiment Task

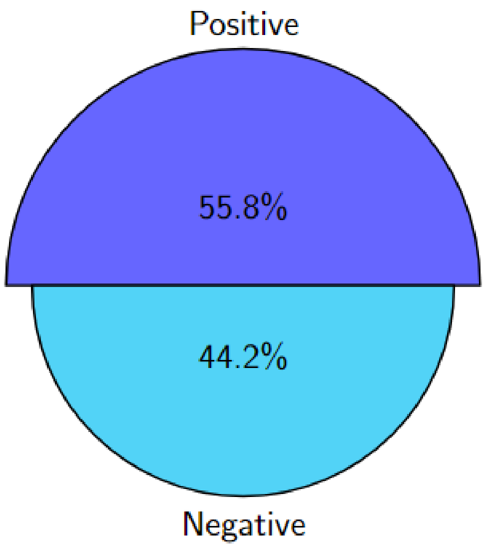


- Includes 15K Romanian reviews collected from one of the largest e-commerce platforms.
- The original split – 11K train, 1K validation, and 3K test will be used in this research.

02 Datasets Establishment

SST-2 Dataset

Sentiment Task

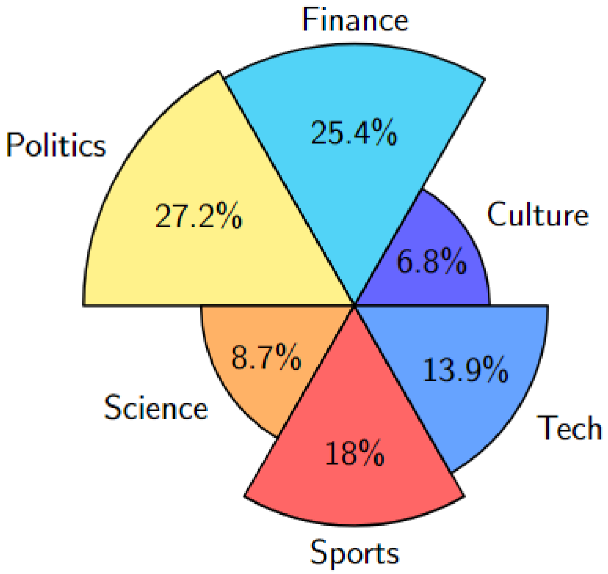


- Selected ~70K English sentences extracted from movie reviews.
- The original split – 96% train, 1.3% validation, and 2.6% test will be used in this research.
- It can be observed that there are various domains, not only in terms of language but also regarding the source domain of the datasets.

03 Datasets Establishment

MoRoCo Dataset

News Categorization Task

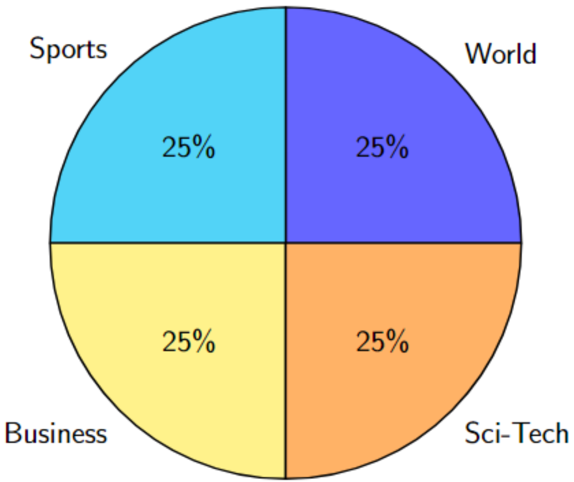


- Selected ~33.5K Romanian and Moldovian samples of text collected from the news domain.
- The original split – ~65% train, 17.6% validation, and 17.6% test will be used in this research.

03 Datasets Establishment

Ag News Dataset

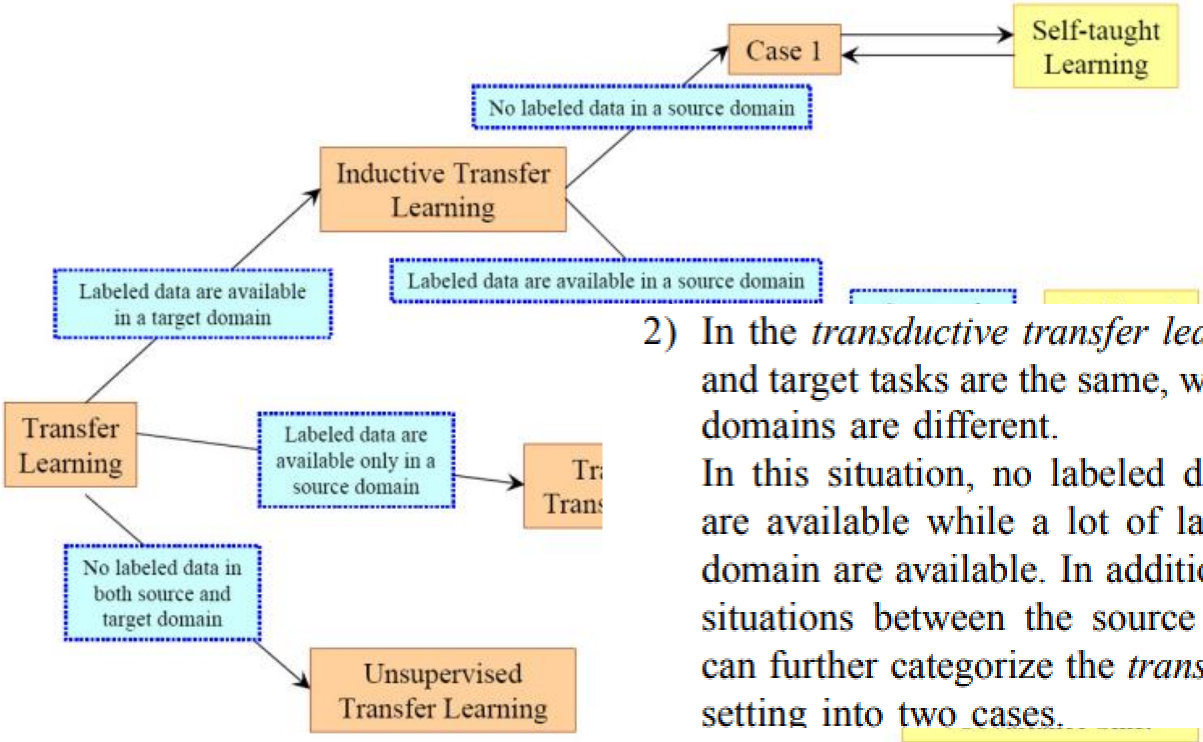
News Categorization Task



- Vast collection of English news articles, comprising over 1 million examples
- Only a subset of 120K examples for training and 7.6K will be used.

The Method

Problem definition: About Transfer Learning



2) In the *transductive transfer learning* setting, the source and target tasks are the same, while the source and target domains are different.
In this situation, no labeled data in the target domain are available while a lot of labeled data in the source domain are available. In addition, according to different situations between the source and target domains, we can further categorize the *transductive transfer learning* setting into two cases.

Problem definition: About Transfer Learning

Cross-Lingual Text Categorization

Nuria Bel[1], Cornelis H.A. Koster[2], and Marta Villegas[1]

[1] Grup d'Investigació en Lingüística Computacional Universitat de Barcelona, 08028 - Barcelona, Spain. {nuria,tona}@gilc.ub.es
[2] Computer Science Dept., University of Nijmegen, Toernooiveld 1, 6525ED Nijmegen, The Netherlands. kees@cs.kun.nl

First Domain Adaptation Method

Abstract

Fine-tuning is known to improve NLP models by adapting an initial model trained on more plentiful but less domain-salient examples to data in a target domain. Such domain adaptation is typically done using one stage of fine-tuning. We demonstrate that gradually fine-tuning in a multi-stage process can yield substantial further gains and can be applied without modifying the model or learning objective.

1 Introduction

Domain adaptation is a technique for practical applications in which one wants to learn a model for a task in a particular domain with too few instances

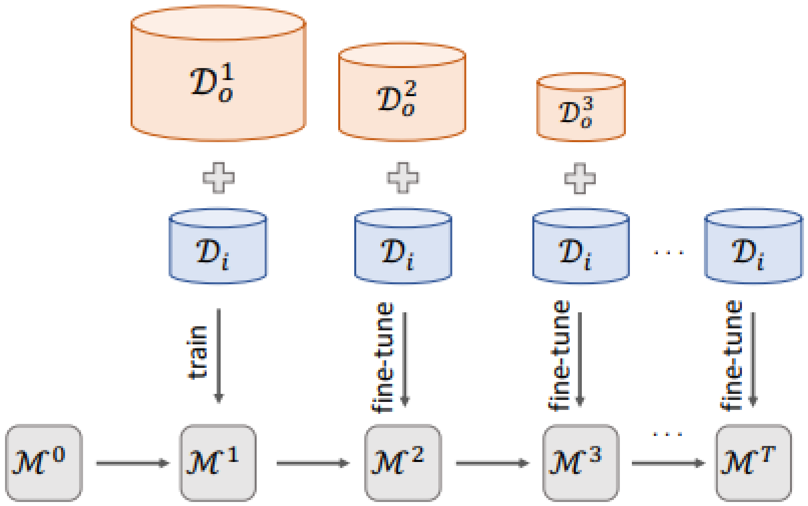
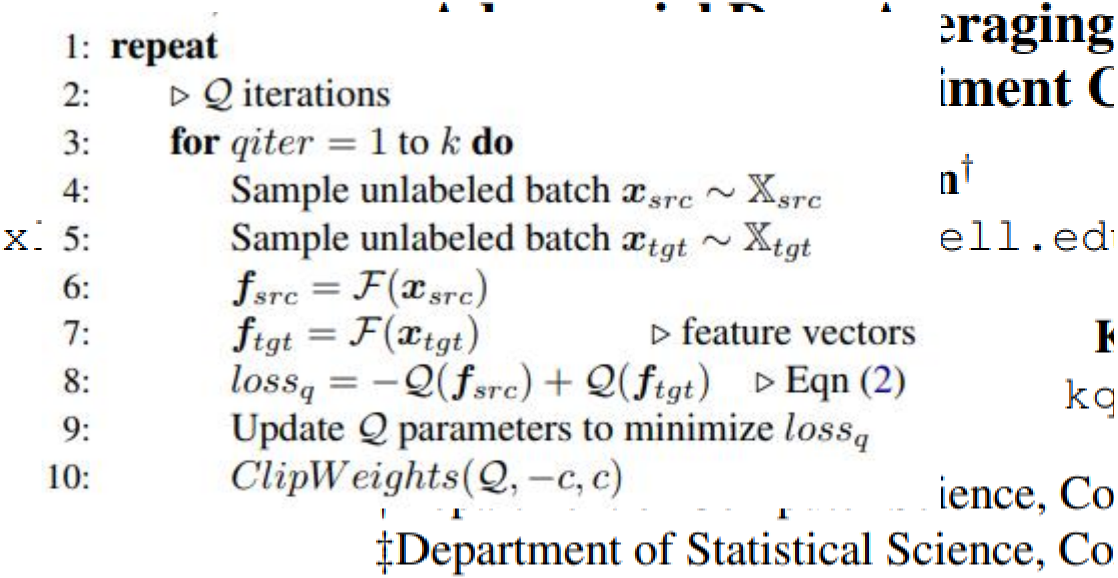


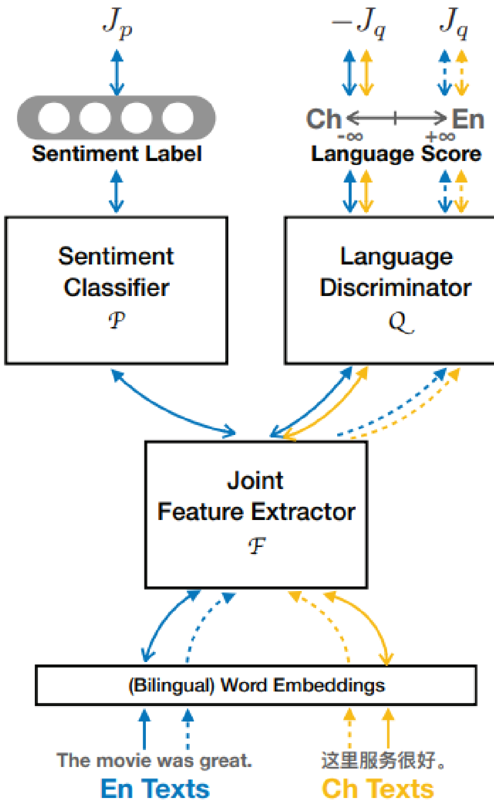
Figure 1: Stages of gradual fine-tuning: 1) Train the model, $\mathcal{M}$, on a mixture of in-domain data, $\mathcal{D}_i$, and out-of-domain data, $\mathcal{D}_o$; 2) iteratively fine-tune on mixed domain data with decreasing amounts of out-of-domain data; 3) fine-tune on only in-domain data.

Adversarial Deep Averaging Networks



repeat
  ▷ $\mathcal{Q}$ iterations
  for $qiter = 1$ to k **do**
    Sample unlabeled batch $\boldsymbol{x}_{src} \sim \mathbb{X}_{src}$
    Sample unlabeled batch $\boldsymbol{x}_{tgt} \sim \mathbb{X}_{tgt}$
    $\boldsymbol{f}_{src} = \mathcal{F}(\boldsymbol{x}_{src})$
    $\boldsymbol{f}_{tgt} = \mathcal{F}(\boldsymbol{x}_{tgt})$ ▷ feature vectors
    $loss_q = -\mathcal{Q}(\boldsymbol{f}_{src}) + \mathcal{Q}(\boldsymbol{f}_{tgt})$ ▷ Eqn (2)
    Update $\mathcal{Q}$ parameters to minimize $loss_q$
    $ClipWeights(\mathcal{Q}, -c, c)$

‡Department of Statistical Science, Co



Born-Again Multi-Task Networks

BAM! Born-Again Multi-Task Networks for Natural Language Understanding

Kevin Clark[†] **Minh-Thang Luong**[‡] **Urvashi Khandelwal**[†]
Christopher D. Manning[†] **Quoc V. Le**[‡]
[†]Computer Science Department, Stanford University
[‡] Google Brain
{kevclark,urvashik,manning}@cs.stanford.edu
{thangluong,qvl}@google.com

Abstract

It can be challenging to train multi-task neural networks that outperform or even match their single-task counterparts. To help address this, we propose using knowledge distillation where single-task models teach a multi-task model. We enhance this training with teacher annealing, a novel method that gradually transitions the model from distillation to supervised learning, helping the multi-task model surpass its single-task teachers. We evaluate

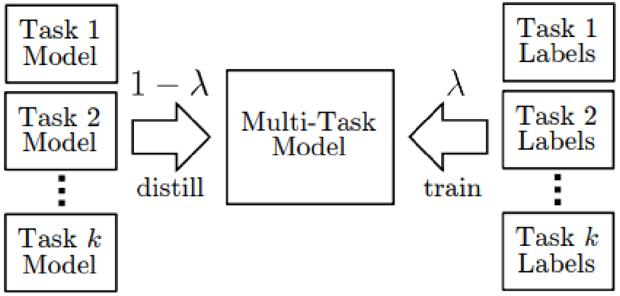


Figure 1: Overview of our method. λ is increased linearly from 0 to 1 over the course of training.

Born-Again Multi-Task Networks

original, benefiting from
involves transferring infoı
model, known as the stude
ability to mimic the teach
fact that the teacher's resı
than a one-hot label. [9] d
the entire technique pres
maintaining an architectu
preserved, while also plac
final model obtained as a

Teacher Annealing. In knowledge distillation, the student is trained to imitate the teacher. This raises the concern that the student may be limited by the teacher's performance and not be able to substantially outperform the teacher. To address this, we propose *teacher annealing*, which mixes the teacher prediction with the gold label during training. Specifically, the term in the summation becomes

$$\ell(\lambda y_\tau^i + (1-\lambda) f_\tau(x_\tau^i, \theta_\tau), f_\tau(x_\tau^i, \theta))$$

where λ is linearly increased from 0 to 1 throughout training. Early in training, the model is mostly distilling to get as useful of a training signal as possible. Towards the end of training, the model is mostly relying on the gold-standard labels so it can learn to surpass its teachers.

f **Knowledge Distillation** [9]
as the teacher, to an initiating
pproach relies on the student's
'he advantage arises from the
l classes, provide more signal
." In this article, I will employ
f teacher annealing [10]. By
nparable learning capacity is
n is learned. I will refer to the
model.

Computer Science & Engineering Department

Thank you

Research advisor:
Șl. Dr. Ing. Dumitru-Clementin Cercel

University Politehnica of Bucharest