Domain_adaptation at SemEval-2025 Task 11: Adversarial Domain Adaptation for Text-based Emotion Recognition

Mikhail Lepekhin
MIPT / Moscow
lepehin.mn@phystech.edu

Serge Sharoff
University of Leeds / UK
s.sharoff@leeds.ac.uk

Abstract

We report our participation in the SemEval-2025 shared task on classification of emotions and describe our solutions using BERT-based models and their modifications. We participate in tracks A and B. We apply and compare base XLM-RoBERTa, Adversarial Domain Adaptation (ADA) on the XLM-RoBERTa with the length of the text as the adversarial feature. As a simple baseline, we also use a Logistic Regression based on tf-idf features. We show that using ADA increases the f1 macro score in low-resource languages and in shorter texts. Besides, we describe our approach to track A where we use ADA with the text language as the confounder. We show that for some languages it helps to improve the f1 score. In all the tracks, we work with the following languages: Russian, Amharic, Algerian Arabic, German, English, Spanish, Hausa, Brasilian Portuguese, Romanian, Ukrainian.

1 Introduction

Non-topical text classification includes a wide range of tasks aimed at predicting a text property that is not connected directly to a text topic. For example, predicting a text style, politeness , difficulty level, the age or the first language of its author, etc. It is applied in many areas such as information retrieval, language teaching, or linguistic research.

(Devlin et al., 2018) introduced BERT – (Bidirectional Encoder Representations from Transformers), an efficient language representation model based on the Transformer architecture (Vaswani et al., 2017). It achieves state-of-the-art results for various NLP tasks, including text classification. XLM-RoBERTa (Conneau et al., 2019) is an improved variant of BERT. It has a similar architecture but uses a bigger and more genre-diverse corpus based on Common Crawl (instead of Wikipedia for the multilingual BERT). Therefore, we choose XLM-RoBERTa as the classifier for the experiments in our research.

One of the most significant problems in text classification is distribution shifts, such as topical shifts, shifts in text length, or the distribution of languages. For example, (Petrenz and Webber, 2010) shows the effect of topical shifts for genre classification. If a topic is more frequent in the training corpus for a given target class, then a classifier tends to predict the target class by the keywords of the topic. This causes numerous unreasonable mistakes in text classification.

One of the algorithms that could be helpful to mitigate topical shifts is Adversarial Domain Adaptation (ADA) (Ganin et al., 2016). It uses an adversarial loss to make the classification features less dependent on the domain of the training data. It supposes training a feature extractor, a domain discriminator, and a target classifier. The feature extractor and target classifier are trained to achieve high accuracy for the classification of the target class and at the same time deceive the domain discriminator to make it impossible to differentiate two domains. In contrast, the domain discriminator intends to classify the text domain correctly.

There was a lot of research on text-based emotion classification in recent years. Some of them use classical ML approaches. For example, (Liu et al., 2023) adjust the Multi-label K-Nearest Neighbors (MLkNN) classifier to allow iterative corrections of the multi-label emotion classification.

In this study, we report our participation in SemEval 2025 task 11 (Muhammad et al., 2025b). We train XLM-RoBERTa base and try to improve its performance with addition of Adversarial Domain Adaptation (Ganin et al., 2016).

2 Related Work

Non-topical text classification is not a new task. For example, numerous attempts have appeared to build a precise classifier of genres based on various architectures from linear discrimination (Karlgren

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 49–53

lang	anger	disgust	fear	joy	sadness	surprise
rus	20.3	10.2	12.2	20.7	15.7	13.3
chn	44.6	15.3	2.7	20	13.4	6.7
deu	29.5	32	9.2	20.8	19.8	6.1
eng	12	0	58.2	24.4	31.7	30.3
esp	24.7	32.8	15.9	32.2	15.5	21.1
ptbr	32.3	3.4	4.9	26.1	14.5	6.9
ukr	4	3.5	7	16.7	13.5	8

Table 1: Tracks A and C. Percentage of positive examples for each emotion in the training data

and Cutting, 1994) to SVM (Sharoff et al., 2010) and recurrent neural networks (Kunilovskaya and Sharoff, 2019).

Most state-of-the art results in the domain of NLP were achieved with transformer-based architectures. (Sun et al., 2019a) gives important advices on how to apply the BERT architecture to the task of text classification. We use the recommended values of learning rate and the number of epochs in our study.

The task of emotion classification is also well-known and widely researched. For example, in (Rasouli and Kiani, 2023) the authors apply a BERT-based transfer learning approach to achieve high accuracy on the short Persian texts. However, their study does not include usage and analysis of the adversarial methods in contrast to ours.

(Zou et al., 2021) modify Adversarial Domain Adaptation (ADA) and present a novel approach for domain adaptation. The methods are applied and compared on the tasks of sentiment analysis and yes-no binary questino answering. Although their results surpasses other techniques compared in their study, the authors mostly work with much longer texts then we do in our study. Regarding the shortness of the texts provided in the SemEval 2025 shared task 11, it cannot be guaranteed that the novel methods are able to significantly overpass the simplier ones.

3 Data Analysis

Before making any experiments, we look at the given data to mention some patterns which could be helpful for building robust classifiers.

All the data we use in our study is provided by the SemEval 2025 shared task 11 organizers (Muhammad et al., 2025a). The dev and test data contain a wide range of languages including the rare ones. For example, it contains the Ethiopian languages (Amharic, Oromo, Somali, and Tigrinya) (Belay et al., 2025).

emotion	intensity			
	0	1	2	3
anger	74.2	14.2	8.1	1.7
disgust	70.1	9.3	6.3	1.5
fear	82.8	7.7	6.0	1.7
joy	77.1	8.6	10.2	2.2
sadness	77.5	10.9	7.2	2.5
surprise	84.5	8.3	4.5	0.9

Table 2: Track B. Distribution of intensity for each emotion in the training data

lang	dev				test			
	mean	p=25	p=50	p=75	mean	p=25	median	p=75
rus	9.1	5	8	12	9.7	5	9	13
amh	20.3	10	17	24	19.9	10	17	24
arq	14.7	10	14	19	14.4	9	13	18
deu	35.1	15.8	27	48	35.4	14.0	26	49
eng	14.9	7	12	21.3	15.8	8.0	13	21
esp	10.4	7	9	14	8.8	5	8	12
hau	13.7	8	12	16	13.5	8	12	16
ptbr	18.6	8	13	22	17	8	14	26
ron	16.5	9.5	14	20.5	17	10	15	21
ukr	10	6	9	13	9.9	6	8	12

Table 3: Track A. The number of words per text by language. Mean, median (or 50-percentile), 25- and 75-percentiles.

Table 1 represents the distribution of emotions across the training datasets for all the languages. It can be seen that the training dataset is sparse as it contains less then 20% positive examples for most pairs (language, emotion). Moreover, Table 1 shows that the languages are quite different in terms of the emotions provided for them in the training dataset.

Table 2 shows that the categories distribution in the train for the track B is even more sprarse than that for tracks A and C.

In Table 3, we compare the languages in terms of the distribution of length. It can be seen that the text length depends crucially on the language it comes from. In addition, it can be concluded that the texts in the training and test datasets are quite short and rarely contain more than 1-3 sentences. It causes an additional challenge to create a reliable text-based classifier.

Table 3 shows that the length distibution for train and test differ statistically noticeably. We perform a t-test and get that for Spanish this difference is statistically significant. Moreover, the languages are different in terms of the length distribution. Is could potentially force the classifiers to learn spurious relations between the text length and the emotion label.

4 Experiments

4.1 Methodology

ADA method belongs to Unsupervised Domain Adaptation (Ramponi and Plank, 2020). It shows promising performance in numerous NLP tasks in recent years (Ganin et al., 2016).

It usually consists of a shared feature extractor $f = G_f(x)$, a label predictor $y = G_y(x)$ and a domain discriminator $d = G_d(x)$. In addition to the standard full supervision learning process in the source domain, a minimax game is designed between the feature extractor f and the domain discriminator d. The domain discriminator d aims to distinguish the domain label between the source and target, while the feature extractor f is trained to deceive the feature discriminator d. This adversarial training process can be formulated as

$$\min_{G_f, G_y} L_y(X_s, Y_s) - \lambda L_f(X_s, X_t),$$

$$\min_{G_d} L_d(X_s, X_t),$$

where L_y is the cross-entropy loss for classification of the target label (in our study, it is the gender of the text author). L_f is the loss of the feature extractor. It denotes the cross-entropy of the classification of the text source. Both L_y and L_f are calculated and optimised with freezing of weights of the domain discriminator. L_d is similar to L_f. However, when it is calculated and optimised, the weights of the feature extractor and the label predictor are frozen.

In our study, we use simple discriminators and feature extractors consisting of single linear layers with an activation.

4.2 Description

We train 3 classifiers: Logistic Regression, XLM-RoBERTa, XLM-RoBERTa with Adversarial Domain Adaptation (ADA). All the experiments were carried out on Google Colab.

We use XLM-RoBERTa with base configuration (12-layer, 768-hidden, 12-heads, 125M parameters, xlm-roberta-base in HuggingFace) as a baseline for all the experiments. In all our experiments, we train the XLM-RoBERTa models for 3 epochs with learning rate=10^{-5}, since these values are proposed in (Sun et al., 2019b).

Logistic Regression is used as a simple baseline. We train it on the tf-idf features corresponding to 1-3 gramms. We take 16000 most relevant n gramms according to the chi2 statistics ($sklearn.feature\_selection.SelectKBest$).

Since the training datasets are small for each language, we train each model on all the languages available in the training dataset simultaneously.

Moreover, given the sparsity of the data, we make upsampling for every Logistic Regression classifier we train. Upsampling is not a perfect solution. However, Logistic Regression tends to errode to a constantly zero-predicting classifier without it. Besides, we train a separate Logistic Regression for each emotion.

lang	dev				test			
	xlm-r	adv len	adv lang	lr	xlm-r	adv len	adv lang	lr
rus	0.789	0.709	**0.806**	0.776	0.796	0.726	**0.818**	0.476
amh	0.337	0.528	0.414	**0.642**	0.367	**0.518**	0.439	0.473
arq	0.141	0.382	0.228	**0.568**	0.105	**0.332**	0.136	0.448
chn	**0.555**	0.448	0.536	0.461	0.569	0.500	**0.590**	0.581
deu	0.519	0.447	0.533	**0.599**	0.536	0.482	**0.578**	0.455
eng	0.528	0.498	0.544	**0.624**	0.497	0.463	**0.548**	0.395
esp	0.733	0.703	0.758	**0.772**	0.717	0.683	**0.746**	0.440
hau	0.198	0.415	0.206	**0.756**	0.197	0.380	0.218	**0.460**
ptbr	0.423	0.409	-	**0.574**	0.437	0.420	-	**0.472**
ukr	0.483	0.438	0.508	**0.598**	0.479	0.452	**0.551**	0.489

Table 4: Track A. The f1 macro score of the XLM-R, XLM-R + ADA on the dev and test datasets

lang	dev		test	
	xlm-r	lr	xlm-r	lr
rus	0.310	**0.428**	**0.485**	0.287
amh	**0.430**	0.360	**0.354**	0.302
arq	**0.295**	**0.295**	0.239	**0.262**
chn	0.517	0.287	**0.545**	0.471
deu	**0.537**	0.475	0.511	0.271
eng	0.276	**0.338**	**0.311**	0.207
esp	0.312	**0.354**	**0.410**	0.211
hau	0.393	**0.433**	**0.466**	0.222
ptbr	0.323	**0.361**	**0.542**	0.347
ukr	**0.494**	0.324	**0.416**	0.300

Table 5: Track B. The f1 macro score of the XLM-R on the dev and test datasets

4.3 Results

Table 4 shows the results of our experiments. We can see that for most big languages (English, Russian, Chinese, Ukrainian, German, Spanish, Portuguese), XLM-R without domain adaptation attains a higher f1 macro score. However, the adversarial domain adaptation technique with the length of the text as the confounder helps to attain much better metrics for small languages. For instance, it can be seen for Amkharian and Hausa.

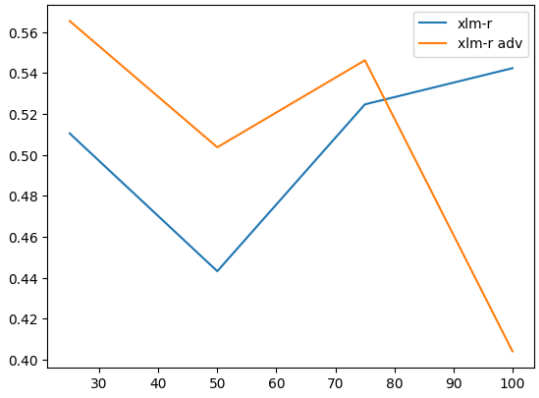

Figure 1: Dependence of the f1 macro score of the base XLM-RoBERTa and the XLM-RoBERTa with ADA on the text length.

After the official deadline for the competition, we also try to use the language as a confounder. Our intuition is that it helps to make the training process more language agnostic. Table 4 shows that this approach manages to beat the base XLM-RoBERTa on most languages for which the training data is available. In track B Table 5, we apply a base XLM-RoBERTa and Logistic Regression based on tf-idf features. We show that the Logistic Regression performs better on most languages on the dev dataset, whilst XLM-RoBERTa attains a higher f1 score for most languages on the test dataset. We suppose it is caused by some sort of distribution shifts between the dev and test datasets.

Besides, the adversarial approach shows Figure 1 significant increase in f1 macro score on the texts of lower length. It shows usability of the adversarial approach and its robustness in case of length distribution shifts.

5 Conclusions and future research

We show that:

1. Using adversarial loss significantly improves the f1 macro score for the low-resource languages

2. Adversarial loss helps to improve the f1 score on the texts with lower length.

3. The metrics for logistic regression are comparable to those for the XLM-RoBERTa models.

Adversarial methods are potentially helpful to achieve higher quality in a wide range of tasks and to combat various distribution shifts, including classification of emotions. However, in order to utilize the whole capacity of the adversarial methods, it would be helpful to use models with a higher number of parameters. For example, best results in the SemEval-2025 Task11 competition (Muhammad et al., 2025b) were achieved using LLMs. However, due to limited computing resources, we did not have the opportunity to fine-tune large language models using the adversarial methods.

Therefore, there is still a room for improvement. In the future, using ADA in conjunction with large language models could make it possible to obtain much more accurate and reliable classifiers. In addition, it might be useful to try more modern competitive domain adaptation methods, such as Energy-based Adversarial Domain Adaptation (EADA).

References

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Chaudhary Vishrav, Guillaume Wenzek, Edouard Grave Francisco Guzman´, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv*, arXiv: 1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Fran¸cois Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research 17 (2016) 1-35*.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *COLING '94: Proc. of the 15th. International Conference on Computational Linguistics*, pages 1071 – 1075, Kyoto, Japan.

Maria Kunilovskaya and Serge Sharoff. 2019. Building functionally similar corpus resources for translation studies. In *Proc RANLP*, Varna.

Xuan Liu, Tianyi Shi, Guohui Zhou, and Mingzhe Liu. 2023. Emotion classification for short texts: an im-

proved multi-label method. *Humanities and Social Sciences Communications*.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Philipp Petrenz and Bonnie Webber. 2010. Stable classification of text genres. *Computational Linguistics*, 34(4):285–293.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. *Coling*.

Mahdi Rasouli and Vahid Kiani. 2023. Investigating shallow and deep learning techniques for emotion classification in short persian texts. *Journal of AI and Data Mining*.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web library of Babel: evaluating genre collections. In *Proc Seventh Language Resources and Evaluation Conference, LREC*, Malta.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019a. How to fine-tune BERT for text classification? *arXiv preprint arXiv:1905.05583*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019b. How to fine-tune bert for text classification? *arXiv preprint arXiv:1905.05583*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Han Zou, Jianfei Yang, and Xiaojian Wu. 2021. Unsupervised energy-based adversarial domain adaptation for cross-domain text classification. *ACL 2021*.