GinGer at SemEval-2025 Task 11: Leveraging Fine-Tuned Transformer Models and LoRA for Sentiment Analysis in Low-Resource Languages

Aylin Naebzadeh*
School of Computer Engineering
Iran University of Science and Technology
Tehran, Iran
aylin.naebzadeh@gmail.com

Fatemeh Askari*
School of Computer Engineering
Sharif University of Technology
Tehran, Iran
fatemeh.askari@ce.sharif.edu

Abstract

Emotion recognition is a crucial task in natural language processing, particularly in the domain of multi-label emotion classification, where a single text can express multiple emotions with varying intensities. In this work, we participated in Task 11, Track A and Track B of the SemEval-2025 competition, focusing on emotion detection in low-resource languages. Our approach leverages transformer-based models combined with parameter-efficient fine-tuning (PEFT) techniques to effectively address the challenges posed by data scarcity. We specifically applied our method to multiple languages and achieved 9^{th} place in the Arabic Algerian track among 40 competing teams. Our results demonstrate the effectiveness of PEFT in improving emotion recognition performance for low-resource languages. The implementation code is publicly available in our GitHub repository[1].

1 Introduction

Sentiment analysis plays a crucial role in understanding human emotions in text, impacting various applications such as customer feedback analysis, social media monitoring, healthcare, and finance. Assigning weights to emotions enhances the precision of sentiment classification, enabling more nuanced decision-making (Jim et al., 2024). With the advancement of deep learning and transformer-based models, sentiment analysis has become more efficient (Cañete et al., 2023; Baziotis et al., 2018; Yu et al., 2018). However, achieving robust accuracy in emotion recognition remains a challenge, especially for low-resource languages, where data scarcity and linguistic diversity hinder model performance.

We focus on categorical emotion classification, where emotions are assigned to discrete categories. Early approaches to textual emotion classification primarily relied on handcrafted features, such as lexicons and rule-based methods (Stone et al., 1966; Strapparava et al., 2004). While modern deep learning models have significantly improved performance (Xu et al., 2020), they are highly dependent on large-scale datasets. When trained on limited data, these models often struggle with overfitting and poor generalization (Tian et al., 2024), making emotion recognition in low-resource settings particularly challenging (Yusuf et al., 2024).

Furthermore, we focus on weighted multi-label text classification, a more complex task where multiple emotions are assigned with varying intensities. While weighting mechanisms enhance emotion modeling, they also come with challenges such as data sparsity, label imbalance, and the difficulty of handling overlapping emotions effectively (Kementchedjhieva and Chalkidis, 2023).

We focus on low-resource languages by leveraging Transformer-based models, evaluating various architectures, including multilingual models. To mitigate overfitting and enhance generalization, we employ parameter-efficient fine-tuning (PEFT) techniques such as LoRA (Low-Rank Adaptation) (Hu et al., 2022), enabling efficient adaptation while maintaining model robustness.

To summarize, we conducted the following experiments on the SemEval 2024 Task 11 dataset:

- Utilizing Transformer-based models to enhance sentiment classification performance.
- Applying PEFT techniques, such as LoRA, to improve efficiency and generalization.
- Assigning density values to each emotion for better sentiment representation.

2 Related Work

Early text classification, including multi-label tasks, relied on traditional machine learning methods

*Authors contributed equally.
[1] https://github.com/AylinNaebzadeh/Text-Based-Emotion-Detection-SemEval-2025

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2028–2037
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

such as Bag-of-Words (BoW) and TF-IDF for feature extraction, using classifiers like Naive Bayes, Support Vector Machines (SVM), and Logistic Regression (Joachims, 1998; Zhang and Zhou, 2005). These approaches represented text as sparse vectors and utilized statistical patterns for classification.

With the rise of deep learning, models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) became popular for text classification (Kim, 2019; Liu et al., 2016). CNNs captured local patterns, while Long Short-Term Memory (LSTM) networks in RNNs excelled in modeling sequential dependencies. Although these methods improved performance by learning dense representations, they struggled with large datasets and long-range dependencies.

The introduction of attention mechanisms and Transformer architectures represented a major advancement (Vaswani et al., 2017; Devlin et al., 2019). Models like BERT and GPT utilized self-attention to capture contextual relationships across documents, surpassing traditional methods in multi-label classification. However, their high computational costs remain a challenge.

To mitigate these issues, Parameter-Efficient Fine-Tuning (PEFT) techniques have emerged, allowing large models to be fine-tuned with reduced computational and memory overhead (Houlsby et al., 2019). Techniques such as LoRA (Low-Rank Adaptation) (Hu et al., 2022), adapters (Houlsby et al., 2019), and prefix tuning (Li and Liang, 2021) facilitate efficient adaptation of pre-trained models to specific tasks, making them more feasible for resource-constrained environments.

3 Task

This SemEval-2025 Task 11: Bridging the Gap in Text-based Emotion Detection (Muhammad et al., 2025a; Belay et al., 2025; Muhammad et al., 2025b) comprises three distinct tracks: Multi-label Emotion Detection (Track A), Emotion Intensity Prediction (Track B), and Cross-lingual Emotion Detection (Track C). Our team participated in the first two tracks. Figure 1 illustrates an overview of the task description.

3.1 Track A

Given a text snippet, the goal is to identify the emotions expressed by the speaker. Specifically, each snippet must be labeled to indicate whether it conveys any of the following emotions: joy, sadness, fear, anger, surprise, or disgust. That is, for each emotion, the snippet is assigned either a positive label (1) if the emotion is present or a negative label (0) if it is absent.

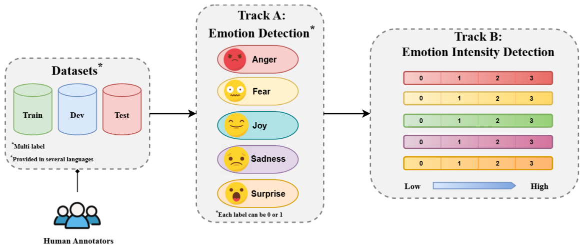


Figure 1: Task Overview for Track A and Track B

For certain languages, such as English, the set of detectable emotions is limited to five—joy, sadness, fear, anger, and surprise—excluding disgust. Table 1 is a sample of the English training data for the first track.

3.2 Track B

For a given text snippet and a specified target emotion, the objective is to predict the intensity level of that emotion.

The possible emotions under consideration include: joy, sadness, fear, anger, surprise, and disgust.

The emotion intensity levels are categorized into the following ordinal classes:

- 0: No emotion present
- 1: Low intensity
- 2: Moderate intensity
- 3: High intensity

Table 2 is a sample of the English training data for the second track.

4 Methodology

Our main focus in the first track was on Afrikaans (AFR), Arabic Algerian (ARQ), Hindi (HIN), and Swedish (SWE) languages. For the second track, we worked on Russian (RUS) and Romanian (RON). To tackle this task, we employ several transformer-based architectures, which are detailed in the Results section. In our experiments, we utilized a consistent set of hyperparameters, including a learning rate of $1e-5$, 100 training epochs, a batch size of 8 for both training and evaluation, and a weight decay of 0.01.

id	text	Joy	Fear	Anger	Sadness	Surprise
eng_train_track1_001	None of us has mentioned the incident since.	0	1	0	1	1
eng_train_track1_002	I was 7 and woke up early, so I went to the basement to watch cartoons.	1	0	0	0	0
eng_train_track1_003	By that point I felt like someone was stabbing my head with a sharp object.	0	1	0	0	0
eng_train_track1_004	watching her leave with dudes drove me crazy.	0	1	1	1	0
eng_train_track1_005	" My eyes widened.	0	1	0	0	1

Table 1: Sample of the English training data for Track A

id	text	Joy	Fear	Anger	Sadness	Surprise
eng_train_track2_001	None of us has mentioned the incident since.	0	1	0	2	1
eng_train_track2_002	I was 7 and woke up early, so I went to the basement to watch cartoons.	1	0	0	0	0
eng_train_track2_003	By that point I felt like someone was stabbing my head with a sharp object.	0	3	0	0	0
eng_train_track2_004	watching her leave with dudes drove me crazy.	0	1	3	1	0
eng_train_track2_005	" My eyes widened.	0	1	0	0	2

Table 2: Sample of the English training data for Track B

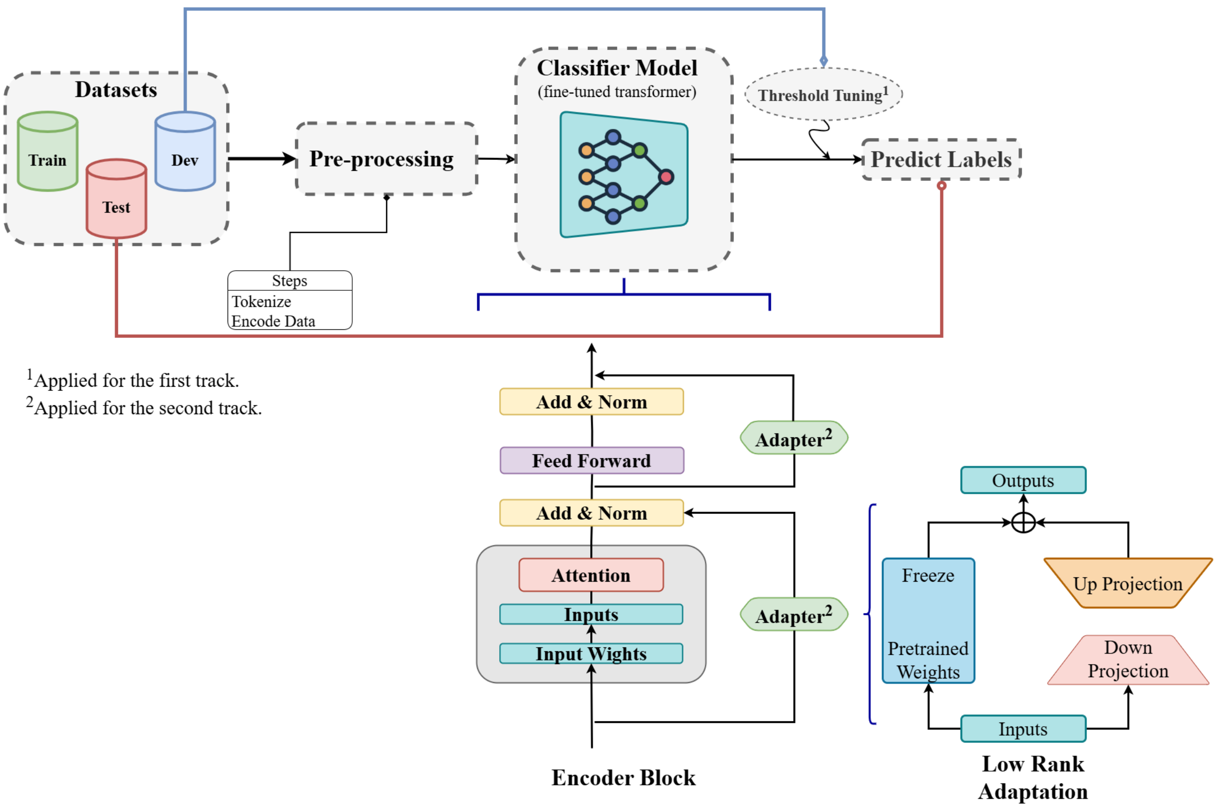


Figure 2: Methodology Overview for Track A and Track B

Figure 2 represents the methodology of our work.

We provide more information about the methodology for each task in separate subsections.

4.1 Track A

For the first track, our approach to multi-label classification involved fine-tuning pretrained transformer models on the training datasets and assessing their performance using the F1 score. During training, we initially set the label threshold in the sigmoid function to 0.3. However, after completing the training process, we applied a threshold tuning strategy to determine the optimal threshold that maximized the F1 score.

4.2 Track B

Our approach to multi-label density prediction (with labels ranging 0–3) combines transformer-based architectures with parameter-efficient fine-tuning strategies.

4.2.1 Parameter-Efficient Fine-Tuning

Since our focus is on low-resource languages, fine-tuning all parameters of large transformer models is computationally expensive and impractical. To mitigate this, we adopt LoRA (Low-Rank Adaptation), a parameter-efficient fine-tuning method that reduces the number of trainable parameters while maintaining performance. LoRA injects trainable low-rank matrices into transformer layers, enabling efficient adaptation to new tasks without modifying the entire model. This approach is particularly beneficial in low-resource scenarios where full fine-tuning would require extensive labeled data and computational resources.

4.2.2 Training Strategy

Loss Function: To optimize our model for the density prediction task, we employ the **Mean Squared Error (MSE)** loss:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

where y_i represents the ground-truth density score (0–3) and $\hat{y}_i$ denotes the predicted value. MSE is chosen for its sensitivity to large deviations, ensuring precise calibration of predicted intensities.

Post-Processing: To enforce annotation guidelines, we apply **floor clipping** to all predictions:

$$\hat{y}_i = \max(0, \min(3, \hat{y}_i))$$

This guarantees outputs remain within the valid range [0, 3].

Evaluation Metric: We measure performance using **Pearson Correlation** for each label:

$$r = \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum (y_i - \bar{y})^2} \sqrt{\sum (\hat{y}_i - \bar{\hat{y}})^2}}$$

This metric evaluates the linear alignment between predictions and ground truth, prioritizing trend consistency over absolute error.

5 Results

The output of confusion matrices and AUC curves on the development datasets are in the appendix section. Performance metrics in Tables 3,4 reveal varying effectiveness of models across languages for emotion detection. The XLM-RoBERTa-Base model (Conneau et al., 2019) scored 0.53 in Afrikaans, while T-XLM-RoBERTa (Barbieri et al., 2022) achieved 0.54. In Hindi, XLM-RoBERTa-Base excelled with 0.84, outperforming T-XLM-RoBERTa (Barbieri et al., 2022) (0.83) and BERT-Multilingual (Devlin et al., 2019) (0.69). For Arabic (Algerian), DiziBERT-Sent. (Abdaoui et al., 2021) scored 0.58, slightly higher than BERT-Multilingual (Devlin et al., 2019) (0.57). In Swedish, BERT-Base-Swedish-Cased-Sent. (Wang et al., 2020) led with 0.72, followed by XLM-RoBERTa-Base (Conneau et al., 2019) (0.71) and T-XLM-RoBERTa (Barbieri et al., 2022) (0.67). Overall, models like XLM-RoBERTa and BERT demonstrate strong performance in emotion detection across multiple languages.

Table 3: Model Performance for Language Emotion on Track A

Language	Model	Micro F1
Afrikaans	XLM-RoBERTa-Base	0.53
	T-XLM-RoBERTa	0.54
Hindi	XLM-RoBERTa-Base	0.84
	T-XLM-RoBERTa	0.83
	BERT-Multilingual	0.69
Arabic (Algerian)	BERT-Multilingual	0.57
	DiziBERT-Sent.	0.58
Swedish	XLM-RoBERTa-Base	0.71
	T-XLM-RoBERTa	0.67
	BERT-Base-Swedish-Cased-Sent.	0.72

Table 4: Model Performance for Language Emotion on Track B

Language	Model	Pearson Corr.
Russian	BERT-Multilingual	0.45
	XLM-RoBERTa	0.83
	T-XLM-RoBERTa	0.74
Romanian	BERT-Multilingual	0.34
	XLM-RoBERTa	0.57
	T-XLM-RoBERTa	0.57

6 Conclusion

Emotion detection and sentiment analysis remain challenging tasks in NLP, particularly for low-resource languages. In this paper, we presented our work and the performance of our models on six low-resource languages in a multi-label classification task using text-based data. Our approach, which leveraged both multilingual and monolingual transformer-based classifiers, demonstrated that these models can achieve notable success. For future work, we aim to explore various hyperparameter settings and investigate the potential of generative models through prompting techniques.

Acknowledgments

We sincerely thank the SemEval-2025 Task 11 organizers for the opportunity to participate in this

valuable experience. We appreciate the chance to work with the provided data, expand our knowledge, and contribute to the field. We also acknowledge the support and encouragement from all those who contributed to our success.

Limitations

Our experiments were constrained by limited hardware resources, preventing us from utilizing models with a higher number of parameters. Additionally, the high cost of certain generative models restricted our ability to explore them further. While some no-cost generative models were available, they often produced outputs in incorrect formats, making them time-consuming to work with for our team.

References

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. Dziribert: a pre-trained language model for the algerian dialect. *arXiv preprint arXiv:2109.12346*.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Christos Baziotis, Nikos Athanasiou, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv:1804.06658*.

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.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. Spanish pre-trained bert model and evaluation data. *arXiv preprint arXiv:2308.02976*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Jamin Rahman Jim, Md Apon Riaz Talukder, Partha Malakar, Md Mohsin Kabir, Kamruddin Nur, and Mohammed Firoz Mridha. 2024. Recent advancements and challenges of nlp-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, page 100059.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

Yova Kementchedjhieva and Ilias Chalkidis. 2023. An exploration of encoder-decoder approaches to multi-label classification for legal and biomedical text. *arXiv preprint arXiv:2305.05627*.

Yoon Kim. 2019. Convolutional neural networks for sentence classification. arxiv 2014. *arXiv preprint arXiv:1408.5882*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

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. *Preprint*, arXiv:2502.11926.

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.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, page 40. Lisbon, Portugal.

Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning, and Prayag Tiwari. 2024. A survey on few-shot class-incremental learning. *Neural Networks*, 169:307–324.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

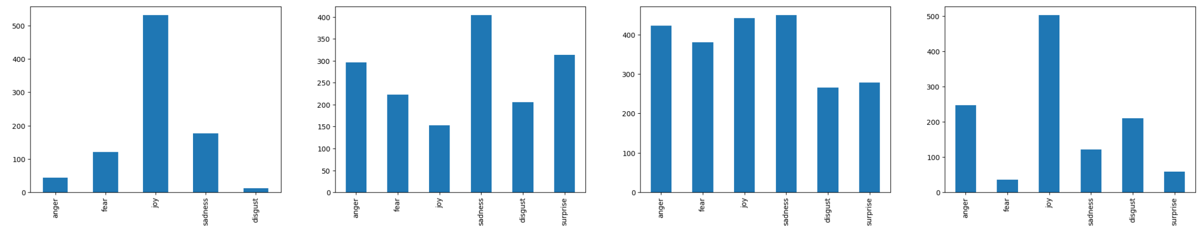
Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Peng Xu, Zihan Liu, Genta Indra Winata, Zhaojiang Lin, and Pascale Fung. 2020. Emograph: Capturing emotion correlations using graph networks. *arXiv preprint arXiv:2008.09378*.

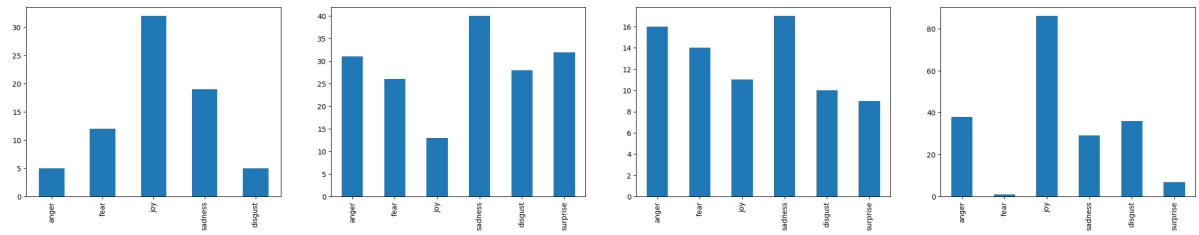
Jianfei Yu, Luis Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. ACL.

Aliyu Yusuf, Aliza Sarlan, Kamaluddeen Usman Danyaro, Abdullahi Sani BA Rahman, and Mujaheed Abdullahi. 2024. Sentiment analysis in low-resource settings: a comprehensive review of approaches, languages, and data sources. *IEEE Access*.

Min-Ling Zhang and Zhi-Hua Zhou. 2005. A k-nearest neighbor based algorithm for multi-label classification. In *2005 IEEE international conference on granular computing*, volume 2, pages 718–721. IEEE.

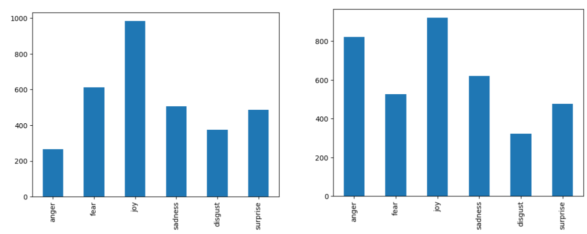


(a) AFR - Train (b) ARQ - Train (c) HIN - Train (d) SWE - Train

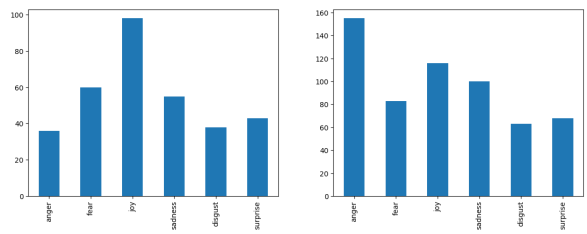


(e) AFR - Dev (f) ARQ - Dev (g) HIN - Dev (h) SWE - Dev

Figure 3: Label Distribution for Train and Dev Datasets per Language in Track A

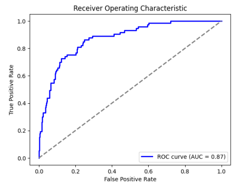


(a) RON - Train (b) RUS - Train

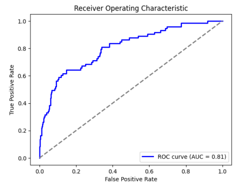


(c) RON - Dev (d) RUS - Dev

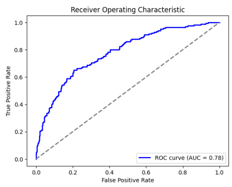
Figure 4: Label Distribution for Train and Dev Datasets per Language in Track B



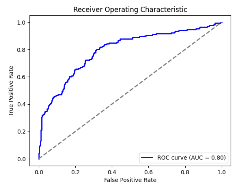
(a) AFR - Twitter XLM RoBERTa Sentiment



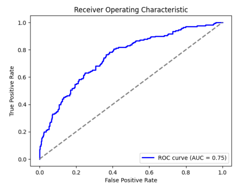
(b) AFR - XLM RoBERTa Base



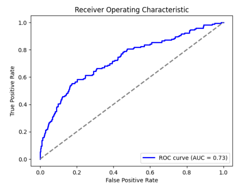
(c) ARQ - AraBERT Base Algerian



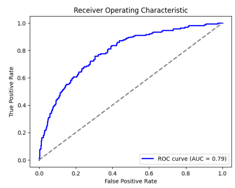
(d) ARQ - AraBERT Large Algerian



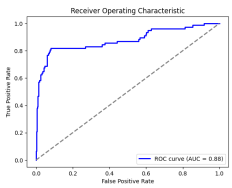
(e) ARQ - AraBERT Medium Algerian



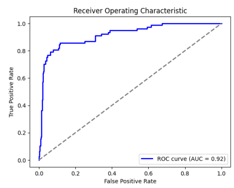
(f) ARQ - AraBERT Mini Algerian



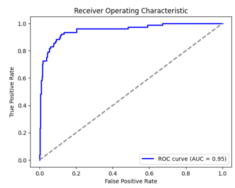
(g) ARQ - Dziri BERT Sentiment



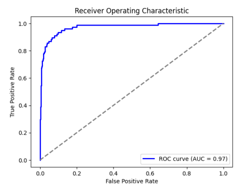
(h) HIN - BERT Base Multilingual Cased



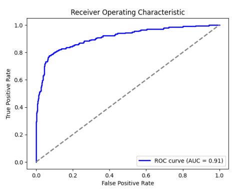
(i) HIN - Multilingual MiniLM L12 H384



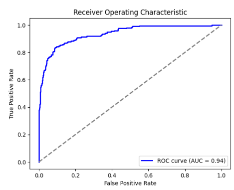
(j) HIN - Twitter XLM RoBERTa Sentiment



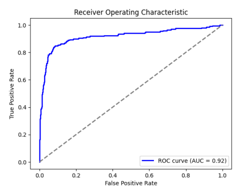
(k) HIN - XLM RoBERTa Base



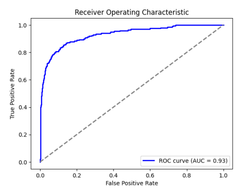
(l) SWE - BERT Base Multilingual Cased



(m) SWE - BERT Base Swedish Sentiment

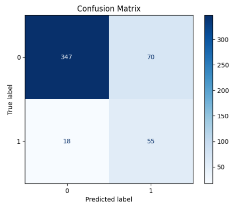


(n) SWE - XLM RoBERTa Base

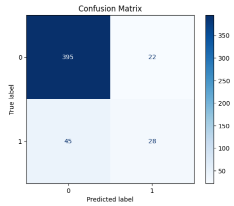


(o) SWE - Twitter XLM RoBERTa Sentiment

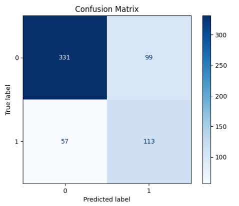
Figure 5: AUC Curves for Models in Different Languages on Dev Datasets



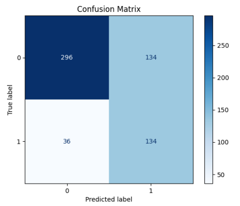
(a) AFR - Twitter XLM RoBERTa Sentiment



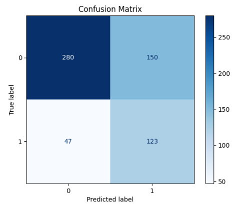
(b) AFR - XLM RoBERTa Base



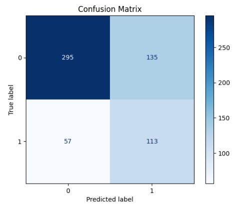
(c) ARQ - AraBERT Base Algerian



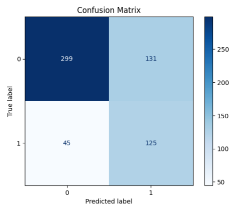
(d) ARQ - AraBERT Large Algerian



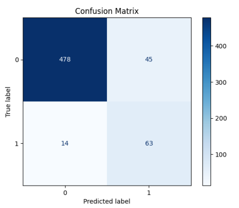
(e) ARQ - AraBERT Medium Algerian



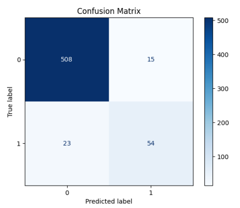
(f) ARQ - AraBERT Mini Algerian



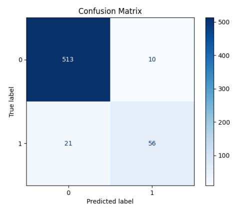
(g) ARQ - Dziri BERT Sentiment



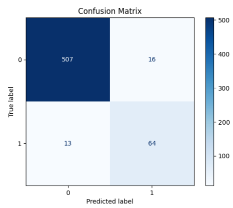
(h) HIN - BERT Base Multilingual Cased



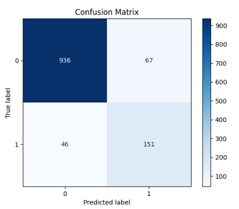
(i) HIN - Multilingual MiniLM L12 H384



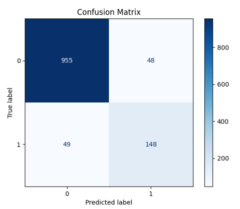
(j) HIN - Twitter XLM RoBERTa Sentiment



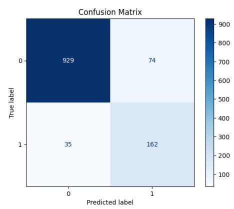
(k) HIN - XLM RoBERTa Base



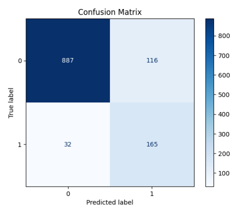
(l) SWE - BERT Base Multilingual Cased



(m) SWE - BERT Base Swedish Sentiment

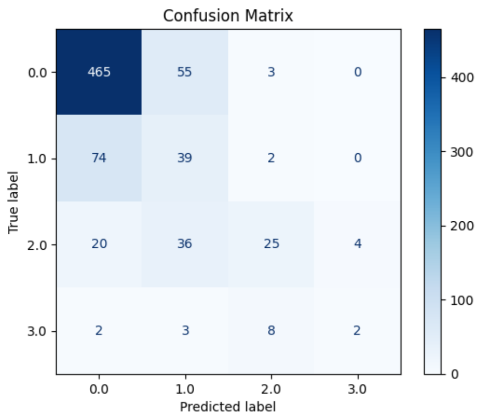


(n) SWE - XLM RoBERTa Base

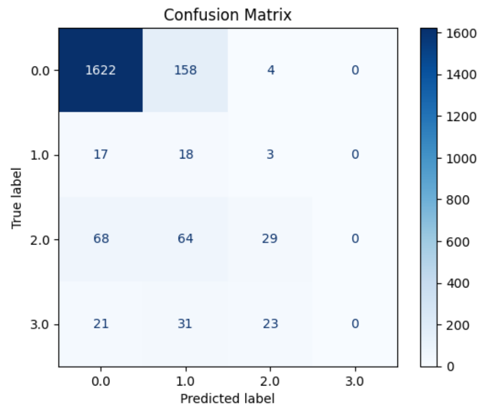


(o) SWE - Twitter XLM RoBERTa Sentiment

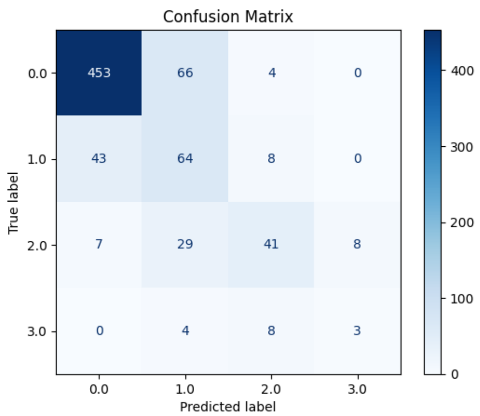
Figure 6: Confusion Matrices for Models in Different Languages on Dev Datasets in Track A



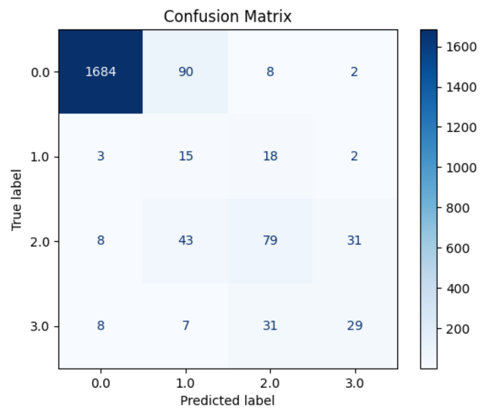
(a) RON - BERT Base Multilingual Cased



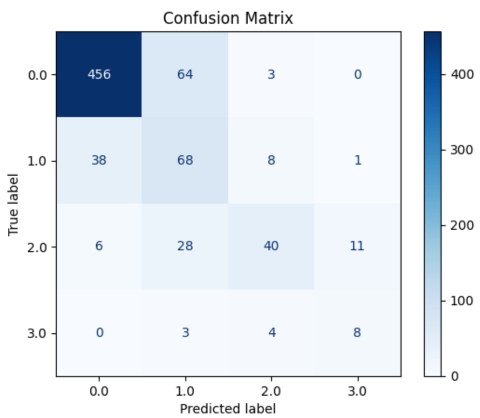
(b) RUS - BERT Base Multilingual Cased



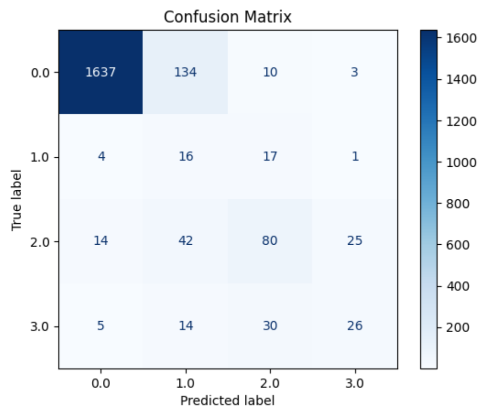
(c) RON - Twitter XLM RoBERTa Sentiment



(d) RUS - Twitter XLM RoBERTa Sentiment



(e) RON - XLM RoBERTa Base



(f) RUS - XLM RoBERTa Base

Figure 7: Confusion Matrices for Models in Different Languages on Dev Datasets in Track B