CSECU-Learners at SemEval-2025 Task 11: Multilingual Emotion Recognition and Intensity Prediction with Language-tuned Transformers and Multi-sample Dropout

Monir Ahmad, Muhammad Anwarul Azim, and Abu Nowshed Chy
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
ahmad.csecu@gmail.com, {azim, nowshed}@cu.ac.bd

Abstract

In today's digital era, individuals convey their feelings, viewpoints, and perspectives across various platforms in nuanced and intricate ways. At times, these expressions can be challenging to articulate and interpret. Emotion recognition aims to identify the most relevant emotions in a text that accurately represent the author's psychological state. Despite its substantial impact on natural language processing (NLP), this task has primarily been researched only in high-resource languages. To bridge this gap, SemEval-2025 Task 11 introduces a multilingual emotion recognition challenge encompassing 32 languages, promoting broader linguistic inclusivity in emotion recognition. This paper presents our participation in this task, where we introduce a language-specific fine-tuned transformer-based system for emotion recognition and emotion intensity prediction. To enhance generalization, we incorporate a multi-sample dropout strategy. Our approach is evaluated across 11 languages, and experimental results demonstrate its competitive performance, achieving top-tier results in certain languages.

1 Introduction

Understanding emotions expressed in a text has gained significant attention in natural language processing (NLP) due to its wide-ranging applications in sentiment analysis, mental health monitoring, and human-computer interaction (Tao and Fang, 2020; Saffar et al., 2023). While sentiment analysis primarily focuses on classifying text into positive, negative, or neutral categories, emotion classification provides a more granular understanding by identifying specific emotions such as joy, sadness, anger, and fear (Mohammad et al., 2018; Ameer et al., 2023).

However, though there are several research on emotion detection in mid- to high-resource languages such as English, Arabic, and Spanish (Mohammad et al., 2018; Saravia et al., 2018; Kumar et al., 2022), very few emotion recognition jobs are done in low-resource languages such as Afrikaans, Hausa, and Romanian (Muhammad et al., 2025a). To bridge this major research gap in emotion recognition, (Muhammad et al., 2025b) introduces a task in SemEval-2025. The task consists of three different tracks. Track A is classifying emotion in a sentence which is structured as a multi-label classification task. Except for English and Afrikaans, sentences in all other languages are required to be classified into six different emotions such as "anger", "fear", "joy", "sadness", "surprise", and "disgust". The "disgust" and "surprise" emotion classes are absent for the English and Afrikaans languages respectively. When a sentence doesn't fall into any of the emotion classes it is categorized as the no emotion instance. Track B is to predict the degree of intensity of each recognized emotion. Track C is to predict the perceived emotion labels of a new text instance in a different target language given a labeled training set in one of the support languages. Among the three tracks, we have participated in the first two. To demonstrate a clear view of the task definition, we articulate an example in Table 1 for the English language.

Sentence	Track A	Track B
I can't believe it! I won the scholarship! This is amazing!	[0 0 1 0 1]	[0 0 3 0 3]

Table 1: Example of Track A and Track B for SemEval-2025 Task 11. In Track A, the values 0 and 1 represent the absence and presence of a specific emotion, respectively. In Track B, the intensity of an emotion is indicated on a scale from 0 to 3, with higher values signifying greater emotional intensity. The classes are presented in the same order as mentioned in the above description.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1332–1341
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

To address the challenges of multilingual and multi-label emotion recognition, as well as emotion intensity prediction, we propose a system in this paper. Our system leverages language-specific transformers to extract contextualized features for a sentence. We utilize a multi-sample dropout strategy for better generalization in our system.

The remaining parts of this paper are organized as follows: Section 2 introduces our proposed system for emotion recognition and emotion intensity prediction. Section 3 details our experimental settings and evaluation. Section 4 offers insightful discussion. Finally, we conclude our paper and suggest potential avenues for future research in Section 5.

2 System Overview

This section provides an overview of our proposed system for SemEval-2025 Task 11: Bridging the Gap in Text-based Emotion Detection. The competition consists of three separate tracks, and we have participated in the first two. Track A focuses on detecting emotions in textual data across multiple languages, while Track B involves predicting emotion intensity. We have participated across 11 languages for both tracks, as summarized in Table 2. Figure 1 presents a high-level illustration of our proposed system.

Given an input sentence, our system first encodes it with a language-tuned transformer (Vaswani et al., 2017). In addition to the contextual embedding from the transformer, we later use a multi-sample dropout (Inoue, 2019; Aziz et al., 2023) procedure to improve the generalization ability of the system. To obtain final logits (unnormalized scores), we fuse the logits from different dropout samples. Finally, we normalize the logits with sigmoid function (Han and Moraga, 1995) and predict with global thresholding.

2.1 Transformer Models

Unlike conventional sequence-based architectures such as LSTM (Schuster and Paliwal, 1997) and CNN (Goodfellow et al., 2016), transformer models effectively capture long-range dependencies within a sequence. Leveraging multi-head attention and positional embeddings enhances token interactions and contextual understanding. We finetune multiple transformer models across various languages to extract contextualized text representations, as illustrated in Table 2.

Language	**Transformer Model**
Amharic (amh)	Davlan/xlm-roberta-base-finetuned-amharic
Algerian Arabic (arq)	Davlan/xlm-roberta-base-finetuned-arabic
Mandarin Chinese (chn)	google-bert/bert-base-chinese
German (deu)	dbmdz/bert-base-german-uncased
English (eng)	Emanuel/twitter-emotion-deberta-v3-base
Spanish (Latin American) (esp)	bertin-project/bertin-roberta-base-spanish (de la Rosa et al., 2022)
Hausa (hau)	Davlan/bert-base-multilingual-cased-finetuned-hausa
Portuguese (Brazilian) (ptbr)	eduagarcia/RoBERTaLexPT-base (Garcia et al., 2024)
Romanian (ron)	readerbench/RoBERT-base (Masala et al., 2020)
Russian (rus)	seara/rubert-base-cased-russian-emotion-detection-cedr
Ukrainian (ukr)	youscan/ukr-roberta-base

Table 2: Language-specific transformers used in our proposed system. The URLs of the Hugging Face models are provided in Table 7 in Appendix A.

2.2 Multi-sample Dropout

In deep neural networks, dropout is an efficient regularization strategy for better generalization (Srivastava et al., 2014). It randomly drops a portion of neurons from the network to prevent dependency among them and hence reduce over-fitting on the training data. As a result, the trained model shows better performance on unseen data. To further enhance the generalization and fast training, (Inoue, 2019) proposed the multi-sample dropout technique. In contrast to the original dropout, features are fed into multiple samples of different dropout masks. Then the output goes to fully connected layers of shared weights. The resulting logits are then used for loss calculation. The final loss is estimated

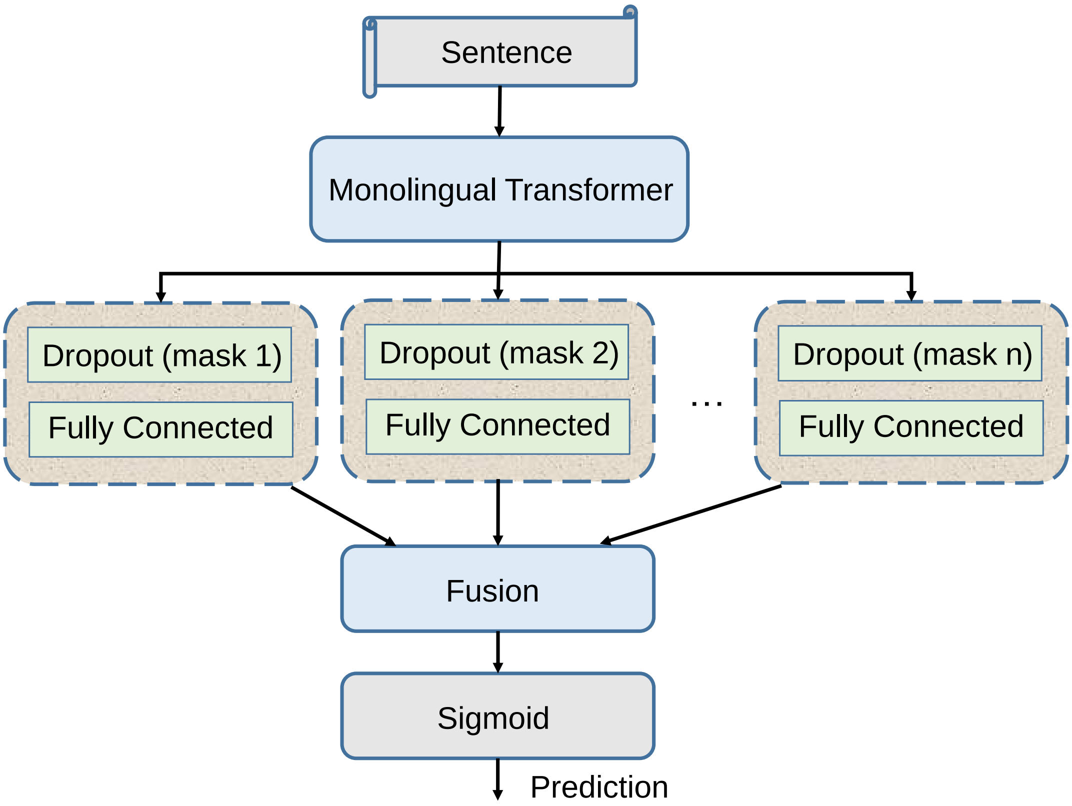

Figure 1: Overview diagram of our proposed system for SemEval-2025 Task 11: Bridging the Gap in Text-based Emotion Detection.

by averaging the observed losses across different samples to achieve a single representation of an input. We utilize a three-sample dropout technique in our proposed system.

2.3 Emotion Classification

During inference, let us get $Logit_1$, $Logit_2$, ... $Logit_n$ from n dropout samples. Then we arithmetic average the logits and pass the output into the sigmoid function (Han and Moraga, 1995) as follows:

$$y = \text{Sigmoid}\left(\frac{\sum_{i=1}^{n} Logit_i}{n}\right) \tag{1}$$

Finally, we predict using thresholding: emotion probabilities y less than the threshold are classified as ‘no emotion’, while those greater than or equal to the threshold are classified as ‘yes’.

2.4 Emotion Intensity Prediction

Track B involves estimating the intensity of emotions for each target class. The intensity levels are classified into four distinct groups: No emotion (0), Low intensity (1), Moderate intensity (2), and High intensity (3). Let p be the probability under which probabilities are predicted to be No emotion (0 intensity) of a class. The remaining probability, $(1 - p)$, is evenly distributed into three segments. Each segment has a length of $\frac{1-p}{3}$, denoted as l. The predicted intensity levels, I, are then determined as follows:

$$I = \begin{cases} \text{No}, & \text{if } y < p, \\ \text{Low}, & \text{else if } p \leq y < p + l, \\ \text{Medium}, & \text{else if } p + l \leq y < p + 2 \times l, \\ \text{High}, & \text{otherwise.} \end{cases} \tag{2}$$

3 Experiments and Evaluation

3.1 Dataset Overview

To demonstrate the effectiveness of participants’ proposed system for emotion classification, the organizers of SemEval-2025 Task 11 have released two benchmark datasets. The BRIGHTER dataset (Muhammad et al., 2025a) covers 28 languages, while EthioEmo (Belay et al., 2025) includes 4 Ethiopian languages, aiming to bridge the gap in text-based emotion recognition. For our participation, we have worked with 10 languages from BRIGHTER and one from EthioEmo (Amharic), covering a total of 11 languages. Figure 2 in Appendix B illustrates the dataset statistics across the train, development, and test sets for these languages. Notably, the development set contains significantly fewer samples compared to the train and test sets for all languages. Both datasets support six emotion classes: “anger”, “fear”, “joy”, “sadness”, “surprise”, and “disgust”. However, the “disgust” class is absent in English, and the “surprise” class is absent in Afrikaans. Additionally, the datasets exhibit a long-tail distribution problem across emotion classes (Muhammad et al., 2025a). During the evaluation stage, we combine the training and development sets to enhance model training and assess its performance on the unseen test set provided in the Codabench competition[1].

3.2 Evaluation Measures

The organizers of SemEval-2025 Task 11 employed various evaluation metrics. The primary metric for Track A and Track C is the macro-average F_1 score. For Track B, which focuses on emotion intensity prediction, the Pearson correlation coefficient is used.

3.3 Parameter Settings

In this section, we outline the configuration details of our proposed system, developed for SemEval-2025 Task 11. We fine-tune multiple transformer models available in Hugging Face (Wolf et al., 2019) for various languages. To ensure reproducibility, we conduct experiments using a T4 GPU on Google Colab (Bisong, 2019), setting the manual seed to 66. We set the classifier learning rate to 0.0001 to facilitate faster convergence. For optimization, we employ the AdamW algorithm (Loshchilov and Hutter, 2017). Additionally, we implement a multi-sample dropout strategy with probabilities ranging from 0.1 to 0.3. The hyperparameter settings and their optimal values are summarized in Table 3. All other parameters remain at their default values.

Hyper-parameters	**Optimal Value**
Batch size	16
Encoder learning rate	3e-5
Number of epochs	9
Max-len	256
Multi-sample dropouts	{0.1, 0.2, 0.3}
Threshold, p in Eq. 2	0.4

Table 3: Hyperparameter settings for our system.

3.4 Results and Analysis

Performance comparison of our proposed system with the baseline (Muhammad et al., 2025b) for Track A and Track B are summarized in Table 4 and 5 respectively. Here, “# Systems” indicates the number of systems reported in the official ranking for a particular language. Following the benchmark of SemEval-2025 task 11, the evaluation is conducted using the primary evaluation metric, macro-average F_1 and Pearson correlation (R) score for track A and track B respectively.

The performance table shows that our system achieves the highest result in the rus language and the lowest result in the ukr language for Track A. For Track B, the highest Pearson correlation is observed for the amh language, while the lowest is for the arq language. Our system performs competitively on the leaderboard in certain languages, achieving the highest Pearson correlation for the amh language among participants. For a detailed comparison of results across participants, we refer to (Muhammad et al., 2025b).

Language	**CSECU-Learners**	**Baseline**	**# Systems**
rus	0.8469 (20th)	0.8377 (25th)	44
esp	0.7689 (20th)	0.7744 (18th)	44
ron	0.7471 (6th)	0.7623 (3rd)	39
eng	0.7381 (29th)	0.7083 (45th)	74
amh	0.7023 (3rd)	0.6383 (15th)	40
hau	0.6735 (9th)	0.5955 (19th)	36
deu	0.6017 (23rd)	0.6423 (16th)	44
chn	0.5999 (16th)	0.5308 (30th)	36
arq	0.5554 (8th)	0.4141 (30th)	36
ptbr	0.5238 (19th)	0.4257 (28th)	37
ukr	0.5062 (23rd)	0.5345 (21st)	36

Table 4: Performance comparison of our proposed system with the baseline for Track A across different languages.

[1] https://www.codabench.org/competitions/3863/

Language	CSECU-Learners	Baseline	# Systems
rus	0.8326 (15th)	0.8766 (9th)	25
esp	0.7145 (11th)	0.7259 (10th)	26
ron	0.6370 (10th)	0.5566 (15th)	22
eng	0.6501 (23rd)	0.6415(24th)	36
amh	0.8558 (1st)	0.5079(11th)	20
hau	0.6562 (6th)	0.2703 (23rd)	23
deu	0.5335 (16th)	0.5621 (13th)	24
chn	0.5711 (10th)	0.4053 (21st)	24
arq	0.4430 (12th)	0.0164 (23rd)	23
ptbr	0.4655 (17th)	0.2974 (20th)	23
ukr	0.4780 (12th)	0.3994 (16th)	21

Table 5: Performance comparison of our proposed system with the baseline for Track B across different languages.

4 Discussion

In this section, we estimate the impact of the multi-sample dropout (MSD) strategy in our CSECU-Learners system. Additionally, we compare our system's results with some state-of-the-art (SOTA) multilingual transformers and large language models (LLMs).

Table 6 presents the impact of the MSD technique on emotion classification and intensity prediction. The results are obtained using tuned thresholds across languages on the test set during the post-evaluation phase. The thresholds are provided in Table 8 in Appendix C. For emotion classification, we observe that the CSECU-Learners system with MSD outperforms its counterpart without MSD in 7 of the 11 languages we participated in. Similarly, the MSD-enabled system achieves better results in 7 languages for intensity prediction. Overall, the MSD strategy contributes an improvement of 0.30% in macro-F$_1$ and 0.31% in Pearson correlation for Track A and Track B, respectively.

Since SemEval-2025 Task 11 focuses on multilingual emotion classification, we compare the performance of our system with several multilingual transformer models. Table 9 in Appendix C presents a comparison between our system and three multilingual transformer-based models: RemBERT (Chung et al., 2020), XLM-R (Conneau et al., 2020), and LaBSE (Feng et al., 2022). This evaluation is conducted using our system's performance during the official evaluation phase. The performance scores for multilingual transformers are taken from BRIGHTER (Muhammad et al., 2025a). The comparison indicates that our system outperforms multilingual transformer-based models in most languages across both tracks.

Large Language Models (LLMs) have recently demonstrated remarkable learning and reasoning capabilities across various downstream tasks. In Table 10 in Appendix C, we present a comparative analysis of several LLMs, including Llama-3.3-70B (Touvron et al., 2023), Qwen2.5-72B (Qwen et al., 2025), and DeepSeek-R1-70B (DeepSeek-AI et al., 2025), alongside our system. The results indicate that our system achieves superior performance over these LLM-based approaches for the majority of languages. This demonstrates the effectiveness of our proposed system in the emotion classification task.

5 Conclusion and Future Direction

This paper presents our proposed system for emotion recognition and intensity prediction. Identifying emotions in a sentence requires more than just superficial analysis; understanding contextual meaning is essential. To address this challenge, we fine-tuned various transformer models across different languages, leveraging their ability to capture contextual embeddings. Additionally, we incorporated a multi-sample dropout strategy to enhance generalization. Experimental results validate the effectiveness of our proposed approach, demonstrating competitive performance in comparison to several existing methods.

In future work, we plan to explore other state-of-the-art transformer architectures and investigate the fusion of multiple transformer models. Since the dataset is imbalanced, we aim to incorporate weighted loss functions to improve learning across all classes.

Limitations

Our proposed system utilizes language-specific transformers, requiring fine-tuning for each language, which can be computationally expensive and time-consuming. Additionally, the performance of the model is influenced by threshold tuning, which may vary across different datasets and may not always generalize well to real-world applications. Furthermore, the system does not address the class imbalance problem in this task, which

Model	rus	esp	ron	eng	amh	hau	deu	chn	arq	ptbr	ukr	Avg.
Track A (Macro F_1)												
CSECU-Learners	**.8493**	**.7693**	.8233	.7379	**.7047**	.6762	**.6028**	.6171	**.5554**	**.5260**	**.5075**	**.6700**
- MSD	.8416	.7623	**.8280**	**.7385**	.6999	**.6810**	.5918	**.6227**	.5519	.5206	.4995	.6670
Track B (Pearson Correlation)												
CSECU-Learners	**.8503**	**.7201**	.7417	**.6549**	**.8553**	.6558	**.5522**	**.5958**	**.4430**	.5030	.4934	**.6423**
- MSD	.8352	.7107	**.7424**	.6491	.8505	**.6700**	.5463	.5852	.4384	**.5085**	**.4950**	.6392

Table 6: Impact of the multi-sample dropout strategy in Track A and Track B. The best performance scores are highlighted in **bold**.

could impact overall performance.

References

Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander Gelbukh. 2023. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534.

Abdul Aziz, Md Akram Hossain, and Abu Nowshed Chy. 2023. Csecu-dsg at semeval-2023 task 4: Fine-tuning deberta transformer model with cross-fold training and multi-sample dropout for human values identification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1988–1994.

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.

Ekaba Bisong. 2019. Google colaboratory. In *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, pages 59–64. Springer.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *Preprint*, arXiv:2010.12821.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Javier de la Rosa, Eduardo G Ponferrada, Paulo Villegas, Pablo González de Prado Salas, Manu Romero, and Marıa Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68:13–23.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Eduardo AS Garcia, Nadia FF Silva, Felipe Siqueira, Hidelberg O Albuquerque, Juliana RS Gomes, Ellen Souza, and Eliomar A Lima. 2024. Robertalexpt: A legal roberta model pretrained with deduplication for portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 374–383.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. Deep learning, volume 1.

Jun Han and Claudio Moraga. 1995. The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International workshop on artificial neural networks*, pages 195–201. Springer.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.

Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems*, 240:108112.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. Robert–a romanian bert model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alieh Hajizadeh Saffar, Tiffany Katharine Mann, and Bahadorreza Ofoghi. 2023. Textual emotion detection in health: Advances and applications. *Journal of Biomedical Informatics*, 137:104258.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3687–3697.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Jie Tao and Xing Fang. 2020. Toward multi-label sentiment analysis: a transfer learning based approach. *Journal of Big Data*, 7(1):1.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Transformer URLs

Table 7 shows the URLs of the Hugging Face Transformers used for each language in our system.

B Dataset Statistics

Figure 2 presents the distribution of the train, development, and test sets for the SemEval-2025 Task 11 dataset. The illustration includes only the languages in which we participated.

C Performance Evaluation

C.1 Optimal Thresholds

The optimal thresholds used in our system for Track A and Track B across different languages on the test set are presented in Table 8. We report the thresholds both with and without the multi-sample dropout (MSD) strategy.

C.2 Performance Comparison

Table 9 presents a comparative analysis between our proposed system and several multilingual transformers. From the various multilingual transformers discussed in BRIGHTER (Muhammad et al., 2025a), we report the top three performing models. Similarly, Table 10 provides a performance analysis of several large language models on this task. All results for multilingual transformers and large language models are taken from BRIGHTER.

Language	Transformer URL
amh	https://huggingface.co/Davlan/xlm-roberta-base-finetuned-amharic
arq	https://huggingface.co/Davlan/xlm-roberta-base-finetuned-arabic
chn	https://huggingface.co/google-bert/bert-base-chinese
deu	https://huggingface.co/dbmdz/bert-base-german-uncased
eng	https://huggingface.co/Emanuel/twitter-emotion-deberta-v3-base
esp	https://huggingface.co/bertin-project/bertin-roberta-base-spanish
hau	https://huggingface.co/Davlan/bert-base-multilingual-cased-finetuned-hausa
ptbr	https://huggingface.co/eduagarcia/RoBERTaLexPT-base
ron	https://huggingface.co/readerbench/RoBERT-base
rus	https://huggingface.co/seara/rubert-base-cased-russian-emotion-detection-cedr
ukr	https://huggingface.co/youscan/ukr-roberta-base

Table 7: URLs of language-specific transformers used in our proposed system.

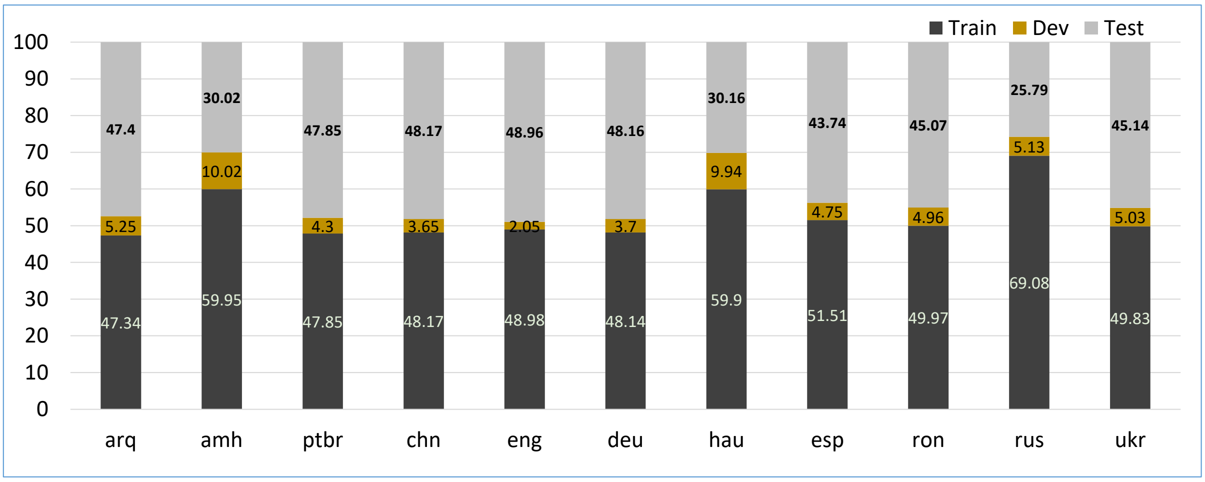

Figure 2: Train, development, and test set percentages for the languages we participated in.

Model	rus	esp	ron	eng	amh	hau	deu	chn	arq	ptbr	ukr
Track A: Emotion Classification											
CSECU-Learners	0.7	0.4	0.3	0.3	0.5	0.3	0.2	0.1	0.2	0.2	0.1
- MSD	0.6	0.5	0.3	0.3	0.4	0.2	0.2	0.1	0.2	0.1	0.1
Track B: Emotion Intensity Prediction											
CSECU-Learners	0.9	0.1	0.2	0.7	0.3	0.1	0.1	0.1	0.2	0.1	0.1
- MSD	0.7	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.3	0.1	0.1

Table 8: Optimal thresholds for Track A and Track B on the test set.

Model	rus	esp	ron	eng	hau	deu	chn	arq	ptbr	ukr
Track A: Emotion Classification										
CSECU-Learners	.8469	.7689	.7471	.7381	.6735	.6017	.5999	.5554	.5238	.5062
RemBERT	.8377	.7744	.7623	.7083	.5955	.6423	.5308	.4141	.4257	.5345
LaBSE	.7562	.7288	.6979	.6424	.5849	.5502	.5347	.4546	.4260	.5007
XLM-R	.7876	.2985	.6521	.6730	.3695	.5537	.5848	.3198	.1540	.1777
Track B: Emotion Intensity Prediction										
CSECU-Learners	.8326	.7145	.6370	.6501	.6562	.5335	.5711	.4430	.4655	.4780
RemBERT	.8766	.7259	.5566	.6415	.2703	.5621	.4053	.0164	.2974	.3994
LaBSE	.6843	.5689	.3557	.3534	.2613	.2893	.2337	.0142	.2062	.1375
XLM-R	.6896	.5572	.3777	.3736	.2468	.3830	.3692	.0089	.1824	.3616

Table 9: Performance comparison between our proposed system and multilingual transformers. The best performance scores in Track A and Track B are highlighted in orange and green, respectively.

Model	rus	esp	ron	eng	hau	deu	chn	arq	ptbr	ukr
Track A: Emotion Classification										
CSECU-Learners	.8469	.7689	.7471	.7381	.6735	.6017	.5999	.5554	.5238	.5062
DeepSeek-R1-70B	.7697	.7329	.6502	.5699	.5191	.5426	.5345	.5087	.5149	.5119
Qwen2.5-72B	.7308	.7233	.6818	.5572	.4379	.5917	.5523	.3778	.5160	.5476
Llama-3.3-70B	.6261	.6127	.7128	.6558	.5091	.5699	.5336	.5575	.4503	.4234
Track B: Emotion Intensity Prediction										
CSECU-Learners	.8326	.7145	.6370	.6501	.6562	.5335	.5711	.4430	.4655	.4780
DeepSeek-R1-70B	.6228	.6074	.5769	.4808	.3885	.5478	.4857	.3637	.4672	.4354
Qwen2.5-72B	.5825	.5111	.5548	.5599	.2700	.4330	.4617	.2954	.3820	.3774
Llama-3.3-70B	.5756	.5164	.4587	.4414	.3916	.5346	.5186	.3629	.4090	.3699

Table 10: Performance comparison between our proposed system and large language models (LLMs). The best performance scores in Track A and Track B are highlighted in orange and green, respectively.