Tue-JMS at SemEval-2025 Task 11: KReLax: An Ensemble-Based Approach for Multilingual Emotion Detection and Addressing Data Imbalance

Jingyu Han
Universität Tübingen

Megan Horikawa
Universität Tübingen

Suvi Lehtosalo
Universität Tübingen

{jingyu.han, megan.horikawa, suvi.lehtosalo}@student.uni-tuebingen.de

Abstract

Emotion detection research has primarily focused on English, leaving a gap for low-resource languages. To address this, we present KReLaX, a multilingual ensemble model for multi-label emotion detection, combining three BERT-based encoders with a weighted voting layer. Within the shared task, our system performed well in multi-label classification, ranking 2nd in Tatar and achieving strong results in Hindi, Russian, Marathi, and Spanish. In emotion intensity classification, we achieved 4th place in Hausa and 5th for Amharic. While our system struggled in the zero-shot track, it achieved 7th place in Indonesian. These results highlight both the potential and the challenges of multilingual emotion detection, emphasizing the need for improved generalization in low-resource settings.

1 Introduction

Emotion detection goes beyond traditional sentiment analysis by interpreting the emotional tone, mood, or psychological state conveyed in an utterance. Emotions are multi-dimensional, context-dependent, and differ across cultures and individuals (Mohammad and Kiritchenko, 2018), making their interpretation in language challenging.

SemEval-2025 Task 11 (Muhammad et al., 2025b) focuses on perceived emotions—determining which emotions most people would attribute to a speaker given a short text. While significant progress has been made in emotion detection in high-resource languages, particularly English (De Bruyne, 2023), there remains a gap in resources and systems for low-resource languages.

To address this, the task encourages the development of multi-label emotion detection systems for underrepresented languages, spanning three tracks:

1. **Track A:** Multi-Label Emotion Detection - predicting multiple emotions per text.
2. **Track B:** Emotion Intensity Classification - predicting the intensity of each emotion.
3. **Track C:** Cross-Lingual Multi-Label Emotion Detection, testing generalization to unseen languages.

In this paper, we present our multilingual ensemble system, KReLaX,[1] developed for all three tracks. Our approach leverages cross-lingual transfer learning (Lin et al., 2019) to improve emotion detection in low-resource languages and applies data augmentation to improve robustness (Wei and Zou, 2019; Dai et al., 2023). Our system is a transformer-based ensemble model that combines multiple multilingual BERT variants with a weighted prediction layer. Prior work has shown that ensemble architectures can help mitigate bias and improve generalization in text classification tasks (Krishnan, 2023; Kumar et al., 2020).

We evaluate our system across all tracks, analyzing the impact of multilinguality, data augmentation, and ensemble learning on performance. Our results show strong performance in multi-label classification (ranking 3rd in Tatar) and emotion intensity classification (6th in Amharic and Hausa). However, zero-shot performance in Track C was challenging, highlighting the need for improved cross-lingual generalization.

2 Task Description

The focus of the task is on perceived emotions: determining which emotions most people would associate with a speaker based on a sentence or short text.

The BRIGHTER collection of emotion recognition datasets was created for the purposes of the task; it contains datasets for 28 languages, including many low-resource languages (Muhammad

[1] Github repository located at https://github.com/HJYnoDebug/KReLaX

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 823–833

et al., 2025a). Four additional languages were drawn from the EthioEmo dataset (Belay et al., 2025). A breakdown of the provided languages and datasets is shown in Table 4.

The training data for the task consists of small texts from various sources in 28 languages for track A, and 11 languages for track B; as track C concerns cross-lingual emotion detection, no training data is provided for this track. Each text is annotated as follows, depending on the task:

- **Track A & Track C (Multi-Label Emotion Detection):** Each emotion — anger, disgust, fear, joy, sadness, and surprise — is labeled as either present ("1") or absent ("0").
- **Track B (Emotion Intensity Classification):** Each emotion is labeled with an intensity score, ranging from 0 to 3.

Each text can have multiple emotion labels, resulting in label imbalance where certain emotions are underrepresented. To address this, we assume independence between emotion labels and approach the task as a series of binary classification problems.

2.1 Evaluation Metric

Track A and Track C are evaluated using the macro F1 score based on our predicted labels and the gold standard labels. The F1 score is the harmonic mean of Precision and Recall for a given class and ensures that model performance is evaluated fairly across all of the labels.

For track B the Pearson Correlation score is used, as it measures the linear relationship between the predicted emotion intensity score and the gold standard scores. This ensures that the models are evaluated based on how well they capture variations in the intensity rather than absolute accuracy.

2.2 Baseline Model

Task organizers provided a fully fine-tuned RemBERT (Chung et al., 2021) model as the baseline. For Tracks A and B, the model was trained and evaluated individually for each language. Class weighting was used in training for track A and C. For track C (Zero-Shot Cross-Lingual Emotion Detection), a family-based leave-one-out approach was used: the target language (i.e. the language being evaluated) was excluded from the training data while retaining other languages from the same family. Baseline results are included in tables 1, 2, and 3.

3 System Overview

Our system implements an ensemble learning approach, where multiple models make predictions, and final predictions are determined via a weighted soft voting layer to determine the final classification. The model architecture diagram is shown in Figure 1.

We fine-tuned three multilingual models in the BERT family - XLM-RoBERTa (Conneau et al., 2019), LaBSE (Wang et al., 2022), and RemBERT (Chung et al., 2021) on sequence classification. Each model was trained on all languages included in track A's training data, excluding Afrikaans due to the difficulty in handling the label alignment (the Afrikaans dataset only included 5 out of 6 emotions — the same was true for English, but we opted to still use the English data due to it being one of the larger datasets provided).

3.1 Stratified Cross-validation

We use stratified K-fold cross-validation (3 folds) to ensure that each fold maintains equal distribution of the labels. The model that performed best on validation data was then selected for final evaluation.

3.2 Class Weighting

To address class imbalance, class weights w_j were computed based on inverse class frequency. First, a scaling factor f_j is derived by dividing the total number of samples N by the product of the number of labels L and the sample count for class j, s_j, with a small smoothing term ϵ to prevent numerical instability:

$$f_j = \frac{N}{L \cdot s_j + \epsilon}, \quad j = 1, 2, \ldots, L. \tag{1}$$

Next, a clipping operation is applied to f_j to ensure that the computed class weights remain within a predefined range, preventing excessively large or small values that could destabilize training. Specifically, the final class weight w_j is constrained within the bounds defined by the lower and upper scaling factors l_b and u_b:

$$w_j = \text{clip}\Big(f_j,\ l_b\, f_j,\ u_b\, f_j\Big). \tag{2}$$

This approach balances the impact of different classes while maintaining numerical stability, thereby improving the robustness and generalization of the model during training.

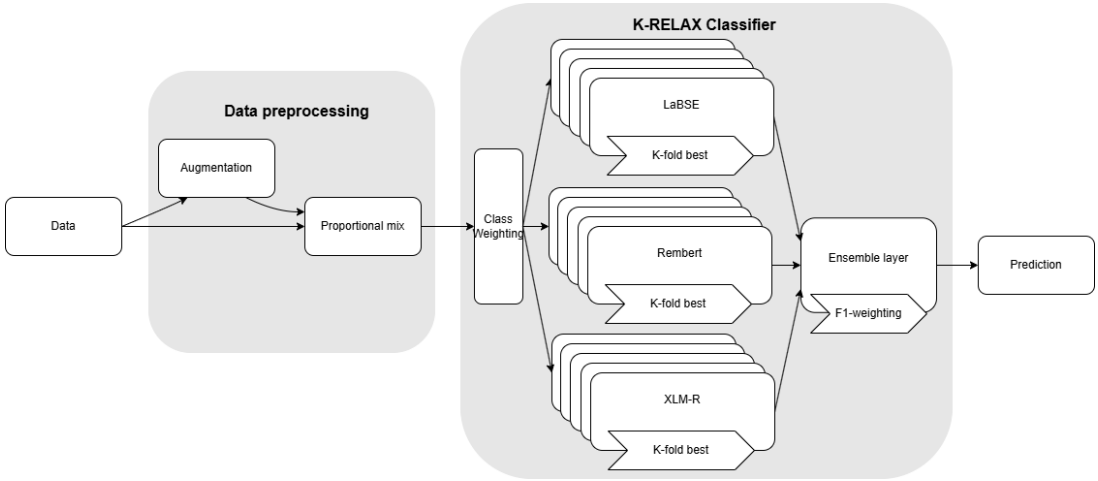

Figure 1: Diagram of the model architecture.

3.3 Classification head for track B

For track B, additional fine-tuning was done using a multi-label multi-class classification head. The model predicts multiple emotion labels per input, where each label has multiple discrete classes.

3.4 SF1-voting

The prediction layer uses a weighted voting system to allow the better performing models more influence over the final classification. The weight for each model is determined by squaring the average F1 score and using it to scale its predictions. By squaring the F1 score we amplify the difference between models in order to give higher performing models more weight and lessen the influence of the poor performing models. The final predictions are obtained by summing the weighted probabilities and normalizing them by the sum of the squared F1 scores as shown in the equation below:

$$\text{final\_probs}(c) = \frac{\sum_{i=1}^{N} (F1_i)^2 \cdot P_i(c)}{\sum_{i=1}^{N} (F1_i)^2} \quad (3)$$

$$\hat{y} = \arg\max_{c} \text{final\_probs}(c) \quad (4)$$

The average F1 score for each individual model is shown in Table 5 in the appendix.

4 Experimental Setup

Our system was fine-tuned on the combined training and development data for track A, excluding Afrikaans. Samples from the augmented data were randomly selected to be included in the training and validation data to balance the classes via stratified sampling, while keeping the proportion of augmented data below 20%. More information on our data augmentation techniques are included in the section below. The training and development datasets were combined and split into 3 proportional folds for cross validation.

We used the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1e-5, a batch size of 16 and a maximum sequence length of 128. Overall, 19.1% of the texts provided for the task were longer than this, with minimal variance across the training, development and test splits, but with large differences between languages (from 0.2% in Spanish to 57.5% in German). Text length seemed to have little effect on accuracy. The model was trained over 20 epochs using 3-fold cross validation with early stopping and a weight decay of .01.

We opted not to use augmented data for track B. This is because substitution and rephrasing may unintentionally affect emotional intensity, and we had no objective way to test for this effect.

4.1 Data Augmentation

We performed data augmentation using methods such as synonym replacement, random swap, back translation, and random deletion as used by Wei and Zou (2019). In the context of this task, random deletion may change the emotion detected in the sentence, and initial tests on back-translation did not lead to any measurable improvements, so we focused on synonym replacement and random swap.

We leveraged a multilingual LLM to generate rephrasings of the training texts as was done in Dai et al. (2023). Suzume-8b (Devine, 2024), a multilingual fine-tuned model based on Llama 3

(Grattafiori et al., 2024) was prompted using 3-shot prompting, an example of which can be found in A.1. The LLM-generated texts were additionally checked for similarity to the original text using the BERT score metric (Zhang et al., 2020). Only generated sentences given a BERT score greater than 70% were used. Augmented data for Amharic, Brazilian Portuguese, English, German, Russian, Somali, Sudanese, and Tigrinya were generated by the LLM.

5 Results

The following sections detail our results in each track. Each model was fine-tuned on the collated training data for track A, which consisted of the languages listed in Table 4, excluding Afrikaans.

5.1 Track A

Our system achieved an average F1 score of .636 across the 24 languages in Track A that we made a submission for, as seen in Table 1. A more detailed breakdown of scores by emotion can be seen in the appendix, in Table 7.

The best overall performance was seen in Hindi, Russian, Marathi, Tatar, and Spanish. Excluding Tatar, where a drop in performance was observed for disgust, our system showed a balanced performance in classification across each emotion. The highest ranking achieved in the task was 2nd place in Tatar.

Across languages, our system was able to consistently detect Joy and Sadness, perhaps due to the larger presence of these emotions across the dataset. Fear, Disgust and Surprise were frequently underrepresented across the dataset, which would support our observations that these emotions were difficult to detect. These emotions may also rely on context and subtle cues that may vary across cultures or languages, making it difficult for the models to generalize cross-linguistically.

Language	**Baseline**	**F1**	Rank
AMH	.6383	.6964	5
ARQ	.4141	.5336	12
ARY	.4716	.5796	5
CHN	.5308	.6033	14
DEU	.6423	.6455	14
ENG	.7083	.6847	57
ESP	.7744	**.7938**	13
HAU	.5955	.6901	5
HIN	.8551	**.8853**	8
IBO	.479	.5297	10
KIN	.4629	.5317	4
MAR	.822	**.8726**	8
ORM	.1263	.5089	12
PCM	.555	.5687	13
PTBR	.4257	.5647	12
PTMZ	.4591	.4782	9
RON	.7623	.7374	15
RUS	.8377	**.8801**	10
SOM	.4593	.4782	8
SUN	.3731	.4389	14
SWE	.5198	.5895	5
TAT	.5394	**.7967**	2
TIR	.4628	.5333	5
UKR	.5345	.6336	8

Table 1: The macro F1 score per language for track A. The top 5 languages are in bold. The left column shows our ranking among other participants in the task.

5.2 Track B

For emotional intensity classification, our system reached an overall average Pearson correlation score of .6724. As shown in Table 2, our model performed best with classifying Russian, Spanish and Amharic, securing a 5th place ranking in Amharic, 4th place in Hausa, and 6th place in Russian. Detailed scores for each emotion are shown in Table 8 in the appendix.

The most consistent performance across emotions was seen in Russian. Among the emotions,

Language	Baseline	Pearson r	Rank
AMH	.5079	**.6716**	5
ARQ	.0164	.4253	13
CHN	.4053	.613	7
DEU	.5621	.6427	8
ENG	.6415	.6653	22
ESP	.7259	**.7386**	9
HAU	.2703	.6698	4
PTBR	.2974	.5598	11
RON	.5566	.654	8
RUS	.8766	**.8863**	6
UKR	.3994	.5608	8

Table 2: Pearson r scores by language in Track B emotion intensity classification.

Joy was the most consistently detected, with strong correlations scores across multiple languages. In contrast, our model struggled to accurately predict the intensity of Surprise, likely due to its low representation in the dataset.

5.3 Track C

Only five of the languages included in the task were not used in training our model; therefore in track C we only submitted results for these five languages, shown in Table 3. Our system faced challenges in zero-shot classification for low-resource languages; the highest-performing language was Indonesian, with a macro F1 score of .5077, ranking 7th. Detailed results are shown in the appendix in Table 9.

The lower performance in isiZulu and isiXhosa could be attributed to their typological differences from the languages included in the training data. In contrast, Indonesian and Javanese may have benefited from the inclusion of Sundanese in the training data, and Afrikaans with German, as they are classified into similar typological language families. Indonesian classification may also have benefited from the large number of loanwords Indonesian has taken from languages such as Hindi, Portuguese, and English (Tadmor, 2009). However, further investigation is needed to determine the extent of language similarity effects on model performance.

6 Conclusion

Using an ensemble method and data augmentation, we developed an emotion classification system that performs well across multiple languages. Our system achieved strong results in Track A, particularly

Language	Baseline	F1 Score	Rank
AFR	.3504	0.3132	10
IND	.3764	**0.5077**	7
JAV	.4638	0.3473	9
XHO	.1273	0.1075	8
ZUL	.1526	0.1309	8

Table 3: The macro F1 score per language for zero-shot emotion classification. The best performing language is in bold.

in Hindi, Russian, Marathi, Spanish, and Tatar, and showed competitive performance in Track B for Amharic, Hausa, and Russian. However, Track B posed greater challenges as the data augmentation methods we used in Track A were not directly applicable to emotion intensity classification. Future work could explore alternative augmentation strategies to preserve intensity information.

Our results in Track C highlight the difficulties of zero-shot classification, particularly for low-resource languages and languages with typological differences from the training data. Performance on unseen languages appears to be influenced by linguistic similarity to training languages, suggesting that further cross-lingual generalization techniques could improve robustness.

Potential future improvements include addressing the label misalignment, experimenting with decoder-based architectures, and refining data augmentation techniques for enhancing both emotion intensity predictions and generalization to unseen languages.

7 Ethical Considerations

With any emotion detection system, there exists a risk that it may be used for harmful purposes, such as governments monitoring social media for negative attitudes in order to target dissidents, or predatory marketing targeting people in a vulnerable emotional state (Mohammad, 2022).

Bias in emotion classification is another challenge, as emotions vary across cultures and languages. Models trained on skewed datasets risk misclassifying or marginalizing underrepresented groups (Janyce Wiebe and Cardie, 2005; Mohammad, 2023; Woensel and Nevil, 2019; De Bruyne, 2023). To mitigate this, transparency in data sources, biases, and limitations are essential to ensuring responsible and fair deployment.

Acknowledgments

We thank the SemEval-2025 Task 11 organizers for their time and efforts in preparing the data and organizing the event so it could run smoothly.

We would also like to thank Çağrı Çöltekin for his encouragement and advice throughout all stages of the task.

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.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. AugGPT: leveraging ChatGPT for text data augmentation. *Preprint*, arXiv:2302.13007.

Luna De Bruyne. 2023. The paradox of multilingual emotion detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 458–466, Toronto, Canada. Association for Computational Linguistics.

Peter Devine. 2024. Tagengo: A multilingual chat dataset. *arXiv preprint arXiv:2405.12612*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing

Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Theresa Wilson Janyce Wiebe and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation 39*.

Anusuya Krishnan. 2023. Optimizing multi-class text classification: A diverse stacking ensemble framework utilizing transformers. *Preprint*, arXiv:2308.11519.

Ayush Kumar, Harsh Agarwal, Keshav Bansal, and Ashutosh Modi. 2020. BAKSA at SemEval-2020 task 9: Bolstering CNN with self-attention for sentiment analysis of code mixed text. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1221–1226, Barcelona (online). International Committee for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Saif Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836, Dubrovnik, Croatia. Association for Computational Linguistics.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saif M. Mohammad. 2022. Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis. *Computational Linguistics*, 48(2):239–278. Place: Cambridge, MA Publisher: MIT Press.

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 Zhangand 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.

Uri Tadmor. 2009. *27. Loanwords in Indonesian*, pages 686–716. De Gruyter Mouton, Berlin, New York.

Weikang Wang, Guanhua Chen, Hanqing Wang, Yue Han, and Yun Chen. 2022. Multilingual sentence transformer as a multilingual word aligner. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2952–2963, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Lieve Van Woensel and Nissy Nevil. 2019. What if your emotions were tracked to spy on you? *European Parliamentary Research Service, PE 634.415.*

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: evaluating text generation with BERT. *Preprint*, arXiv:1904.09675.

A Appendix

A.1 Data Augmentation Prompt

You are a helpful assistant that rephrases text while preserving its original meaning, tone, and style. Ensure the rephrased version is also in the same language and accurately reflects the given emotions. Adjust language to enhance clarity and flow as needed, without altering the message's intent or emphasis. Please output {k} unique rephrased sentences in JSON format. Here are some examples:

```
{examples}
```

Now it is your turn. Here is all the information you need:

```
Emotion(s): {emotion_label}
Language: {lang}
Text to rephrase: {text}
```

A.2 Additional Tables

Language (Code)	Train	Dev	Test
Afrikaans (AFR)	1222	98	1065
Amharic (AMH)	3549	592	1774
Algerian Arabic (ARQ)	901	100	902
Moroccan Arabic (ARY)	1608	267	812
Chinese (CHN)	2642	200	2642
German (DEU)	2603	200	2604
English (ENG)	2768	116	2767
Spanish (Latin American) (ESP)	1996	184	1695
Hausa (HAU)	2145	356	1080
Hindi (HIN)	2556	100	1010
Igbo (IBO)	2880	479	1444
Indonesian (IND)	-	156	851
Javanese (JAV)	-	151	837
Kinyarwanda (KIN)	2451	407	1231
Marathi (MAR)	2415	100	1000
Oromo (ORM)	3442	574	1721
Nigerian-Pidgin (PCM)	3728	620	1870
Portuguese (Brazilian) (PTBR)	2226	200	2226
Portuguese (Mozambican) (PTMZ)	1546	257	776
Romanian (RON)	1241	123	1119
Russian (RUS)	2679 / 2220	199 / 343	1000 / 650
Somali (SOM)	3392	566	1696
Sundanese (SUN)	924	199	926
Swahili (SWA)	3307	551	1656
Swedish (SWE)	1187	200	1188
Tatar (TAT)	1000	200	1000
Tigrinya (TIR)	3681	614	1840
Ukrainian (UKR)	2466	249	2234
Emakhuwa (VMW)	1551	258	777
isiXhosa (XHO)	-	682	1594
Yoruba (YOR)	2992	497	1500
isiZulu (ZUL)	-	875	2047

Table 4: Size of each provided dataset. Languages included in track B are bolded; track B datasets were identical in size to the track A/C sets except for Russian, where the size of the track B set is shown following the slash.

Model	Track A	Track B
XLM-RoBERTa	.6079 ± .0154	.5465 ± .0063
LaBSE	.6302 ± .0100	.5389 ± .0137
RemBERT	.6469 ± .0040	.5584 ± .0226

Table 5: The average macro F1 score (mean ± standard deviation) of the individual models after training.

	Anger	Disgust	Fear	Joy	Sadness	Surprise
Afrikaans	44	12	121	531	177	-
Amharic	1188	1268	109	549	771	151
Brazilian Portuguese	718	75	109	581	322	153
German	768	832	239	541	516	159
English	333	-	1611	674	878	839
Oromo	646	557	123	1091	298	129
Russian	543	273	328	555	421	355
Somali	328	477	305	595	391	179
Sundanese	84	68	47	672	212	226
Tigrinya	547	1311	138	417	588	355
Total	5199	4873	3130	6206	4574	2546

Table 6: Distribution of emotion labels in provided training data.

Language	**Macro F1**	**Anger**	**Disgust**	**Fear**	**Joy**	**Sadness**	**Surprise**
Amharic	.6964	.6626	.7621	.5618	.7535	.7456	.6928
Arabic (Algerian)	.5336	.5529	.4227	.5277	.4324	.6385	.6276
Arabic (Moroccan)	.5796	.5965	.5424	.4815	.7094	.7036	.4444
Chinese	.6033	.8466	.4116	.4000	.8712	.6204	.4702
German	.6455	.7830	.7342	.4504	.7573	.7017	.4463
English	.6847	.5098	–	.8100	.6893	.6970	.7173
Spanish	**.7938**	.7266	.7760	.8359	.8495	.8327	.7420
Hausa	.6901	.5689	.8243	.7750	.6477	.7524	.5724
Hindi	**.8853**	.8452	.8658	.9296	.9062	.8669	.8984
Igbo	.5297	.6316	.4771	.4859	.7506	.6368	.1961
Kinyarwanda	.5317	.4696	.9244	.4314	.6686	.6613	.0351
Marathi	**.8726**	.8283	.8912	.9371	.8134	.8578	.9076
Oromo	.5089	.4832	.5305	.1127	.8242	.3710	.7317
Nigerian Pidgin	.5687	.3386	.7609	.3844	.7162	.6693	.5430
Portuguese (Brazil)	.5647	.7447	.2308	.4804	.7898	.6740	.4688
Portuguese (Mozambique)	.4782	.2941	.2222	.6667	.5319	.6617	.4928
Romanian	.7374	.6018	.7129	.8655	.9589	.7584	.5269
Russian	**.8801**	.8677	.8696	.9524	.9027	.8321	.8560
Somali	.4782	.3565	.3240	.5581	.5959	.6589	.3759
Sundanese	.4389	.2881	.3182	.0952	.9027	.7146	.3146
Swedish	.5895	.7429	.6889	.3556	.9448	.6232	.1818
Tatar	**.7967**	.7280	.6448	.8696	.8603	.8326	.8452
Tigrinya	.5333	.2467	.7154	.3158	.5627	.6000	.7592
Ukrainian	.6336	.4317	.4576	.8296	.7425	.7099	.6306
Mean	.6356	.5894	.6134	.5880	.7576	.7009	.5615

Table 7: Macro F1 scores and emotion classification breakdown for each language in track A. The top 5 languages are in bold.

Language	Average r-score	Anger	Disgust	Fear	Joy	Sadness	Surprise
Amharic	0.6716	0.5232	0.6516	0.6473	0.7816	0.7916	0.6341
Arabic (Algerian)	0.4253	0.435	0.2556	0.4854	0.4974	0.4212	0.457
Chinese (Mandarin)	0.613	0.7373	0.4082	0.5548	0.8762	0.634	0.4675
German	0.6427	0.7383	0.6559	0.464	0.7749	0.7067	0.5165
English	0.6653	0.5557	-	0.6849	0.742	0.712	0.6318
Spanish	0.7386	0.6697	0.6677	0.8095	0.7852	0.8054	0.6941
Hausa	0.6698	0.5417	0.8574	0.7055	0.6688	0.693	0.5523
Portuguese (Brazil)	0.5598	0.6275	0.1697	0.5457	0.7606	0.7203	0.5351
Romanian	0.654	0.5639	0.6413	0.7811	0.9332	0.7124	0.2921
Russian	0.8863	0.864	0.885	0.9489	0.8834	0.8925	0.8439
Ukrainian	0.5608	0.4613	0.2275	0.7652	0.7103	0.6663	0.5339

Table 8: Pearson r-scores by Language and Emotion for track B.

Language	Macro F1	Anger	Disgust	Fear	Joy	Sadness	Surprise
Afrikaans	0.3132	0.1875	0.3478	0.2302	0.3681	0.4324	-
Indonesian	**0.5077**	0.4388	0.3301	0.3689	0.8253	0.7051	0.3781
Javanese	0.3473	0.2443	0.1165	0.0392	0.6643	0.6963	0.3232
isiXhosa	0.1075	0.1096	0.0000	0.0000	0.1621	0.3580	0.0154
isiZulu	0.1309	0.0444	0.0202	0.0278	0.3385	0.3440	0.0105

Table 9: The Macro F1 and emotion scores for each language in track C. The best performing language for Macro F1 is in bold.