SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection

Shamsuddeen Hassan Muhammad[1,2*], **Nedjma Ousidhoum**[3*],
Idris Abdulmumin[4], **Seid Muhie Yimam**[5], **Jan Philip Wahle**[6], **Terry Ruas**[6],
Meriem Beloucif[7], **Christine De Kock**[8], **Tadesse Destaw Belay**[9,10], **Ibrahim Said Ahmad**[11],
Nirmal Surange[12], **Daniela Teodorescu**[13], **David Ifeoluwa Adelani**[14,15,16], **Alham Fikri Aji**[17],
Felermino Ali[18], **Vladimir Araujo**[19], **Abinew Ali Ayele**[5,20], **Oana Ignat**[21],
Alexander Panchenko[22,23], **Yi Zhou**[3], **Saif M. Mohammad**[24]

[1]Imperial College London, [2]Bayero University Kano, [3]Cardiff University, [4]DSFSI, University of Pretoria,
[5]University of Hamburg, [6]University of Göttingen, [7]Uppsala University, [8]University of Melbourne, [9]Instituto Politécnico Nacional,
[10]Wollo University, [11]Northeastern University,[12]IIIT Hyderabad, [13]University of Alberta, [14]MILA, [15]McGill University,
[16]Canada CIFAR AI Chair,[17]MBZUAI,[18]LIACC, FEUP, University of Porto, [19]Sailplane AI, [20]Bahir Dar University,
[21]Santa Clara University, [22]Skoltech, [23]AIRI, [24]National Research Council Canada

Contact: s.muhammad@imperial.ac.uk, OusidhoumN@cardiff.ac.uk

Abstract

We present our shared task on text-based emotion detection, covering more than 30 languages from seven distinct language families. These languages are predominantly low-resource and are spoken across various continents. The data instances are multi-labeled with six emotional classes, with additional datasets in 11 languages annotated for emotion intensity. Participants were asked to predict labels in three tracks: (a) multilabel emotion detection, (b) emotion intensity score detection, and (c) cross-lingual emotion detection.

The task attracted over 700 participants. We received final submissions from more than 200 teams and 93 system description papers. We report baseline results, along with findings on the best-performing systems, the most common approaches, and the most effective methods across different tracks and languages. The datasets for this task are publicly available.

1 Introduction

People use language in diverse and sophisticated ways to express emotions across languages and cultures (Wiebe et al., 2005; Mohammad and Kiritchenko, 2018; Mohammad et al., 2018a). Emotions are also perceived subjectively, even within the same culture or social group. Recognising these emotions is central to language technologies and NLP applications in healthcare, digital humanities, dialogue systems, and beyond (Mohammad et al., 2018b; Saffar et al., 2023). In this work, we use *emotion recognition* to refer to *perceived* emotions, i.e., the emotion most people believe the speaker might have felt based on a sentence or short text snippet.

Despite the linguistic diversity of regions such as Africa and Asia, which together account for more than 4,000 languages, few emotion recognition resources exist for these languages. Prior SemEval shared tasks on emotion recognition have primarily focused on high-resource languages such as English, Spanish, and Arabic (Strapparava and Mihalcea, 2007; Mohammad et al., 2018a; Chatterjee et al., 2019). In this task, we provide participants with new datasets covering more than 30 languages from seven distinct language families, spoken across Africa, Asia, Latin America, North America, and Europe (Muhammad et al., 2025). Our manually annotated emotion recognition datasets, curated in collaboration with local communities, consist of over 100,000 multi-labeled instances drawn from diverse sources, including speeches, social media, news, literature, and reviews. Each instance is labeled by fluent speakers and annotated with six emotion classes: *joy, sadness, anger, fear, surprise, disgust*, and neutral. Additionally, eleven datasets include four emotional intensity levels ranging from 0 to 3 (i.e., absence of emotion to high intensity).

The task consists of three tracks: (a) multilabel emotion detection, (b) emotion intensity detection, and (c) cross-lingual emotion detection. The languages for each track are listed in Figure 1. Each team could submit results for one, two, or all three tracks in one or more languages. Our official evaluation metrics were the average F-score for Tracks A and C and the Pearson correlation coefficient for Track B, which measures how well system-

[*] Equal contribution

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2558–2569
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

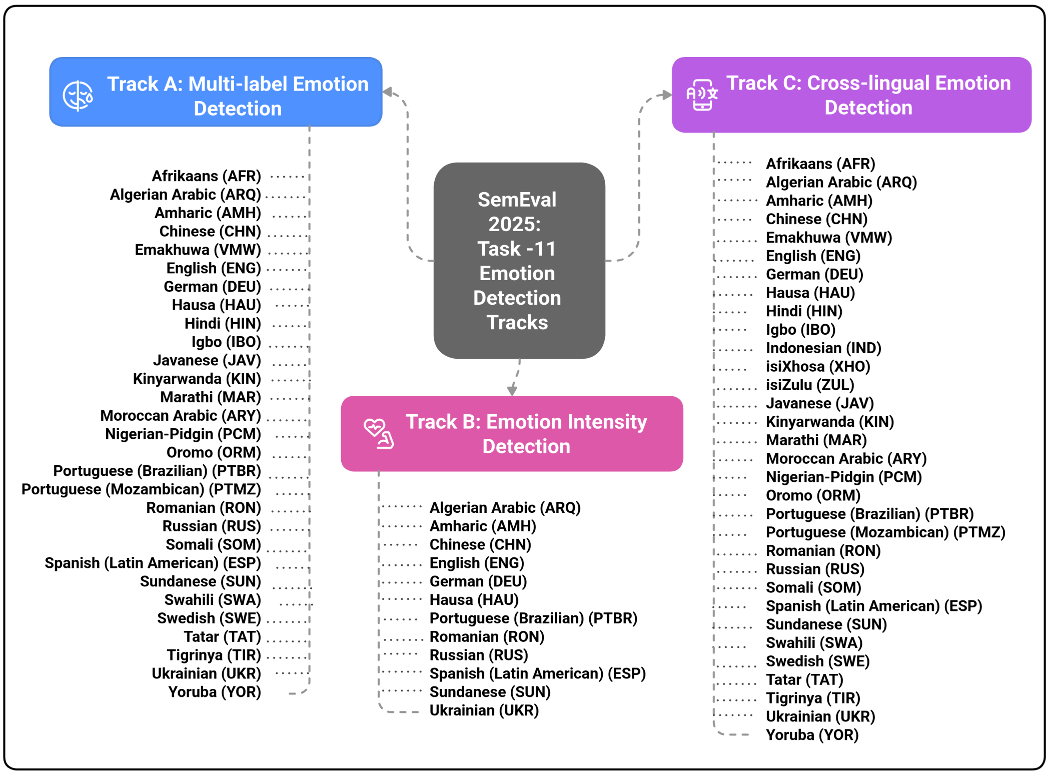

Figure 1: Languages in the three tracks (A, B, and C) of SemEval 2025 Task 11.

predicted intensity scores align with human judgments.

Our task attracted over 700 participants, with 220 final submissions and 93 teams submitting system description papers. Track A (multi-label emotion detection) received the most submissions (114), followed by Track C (cross-lingual emotion detection) with 51, and Track B (emotion intensity detection) with 32. Most teams participated in multiple languages, averaging 11 languages per team. Our task was the most popular competition on CodaBench in 2024.All task details and resources are available on the task's GitHub page.

2 Related Work

NLP work on emotion detection is predominantly Western-centric, with a few exceptions for languages other than English (e.g., Italian (Bianchi et al., 2021), Romanian (Ciobotaru et al., 2022), Indonesian (Saputri et al., 2018), and Bengali (Iqbal et al., 2022)). While multilingual datasets (e.g., (Öhman et al., 2020) and XLM-EMO (Bianchi et al., 2022)) exist, they do not fully capture cultural nuances in emotional expressions due to their reliance on translated data (e.g., XLM-EMO), as emotions are highly contextualized and culture-specific (Havaldar et al., 2023; Mohamed et al., 2024; Hershcovich et al., 2022). Furthermore, most datasets are single-labeled, and to the best of our knowledge, there are no multilingual resources that capture simultaneous emotions and their intensity across various languages.

Additionally, most prior emotion recognition shared tasks have focused on high-resource languages such as English, Spanish, German, and Arabic (Strapparava and Mihalcea, 2007; Mohammad and Bravo-Marquez, 2017; Mohammad et al., 2018b; Chatterjee et al., 2019). In contrast, this shared task covers more than 30 languages, including several low-resource languages.

3 Data

3.1 Data Collection

As our task includes more than 30 different datasets, curated and annotated by fluent speakers, we selected data sources based on: 1) the availability of textual data potentially rich in emotions, and 2) access to annotators. Since finding suitable data is challenging when resources are limited, we typically combine sources. The main textual sources used to build our dataset collection are:

- **Social media posts**: Data collected from var-

Language	**Train**	**Dev**	**Test**	**Total**
Afrikaans (**afr**)	2,107	98	1,065	3,270
Amharic (**amh**)	3,549	592	1,774	5,915
Algerian Arabic (**arq**)	901	100	902	1,903
Moroccan Arabic (**ary**)	1,608	267	812	2,687
Chinese (**chn**)	2,642	200	2,642	5,484
German (**deu**)	2,603	200	2,604	5,407
English (**eng**)	2,768	116	2,767	5,651
Latin American Spanish (**esp**)	1,996	184	1,695	3,875
Hausa (**hau**)	2,145	356	1,080	3,581
Hindi (**hin**)	2,556	100	1,010	3,666
Igbo (**ibo**)	2,880	479	1,444	4,803
Indonesian (**ind**)	–	156	851	1,007
Javanese (**jav**)	–	151	837	988
Kinyarwanda (**kin**)	2,451	407	1,231	4,089
Marathi (**mar**)	2,415	100	1,000	3,515
Nigerian-Pidgin (**pcm**)	3,728	620	1,870	6,218
Oromo (**orm**)	3,442	575	1,721	5,738
Portuguese (Brazilian; **ptbr**)	2,226	200	2,226	4,652
Portuguese (Mozambican; **ptmz**)	1,546	257	776	2,579
Romanian (**ron**)	1,241	123	1,119	2,483
Russian (**rus**)	2,679	199	1,000	3,878
Somali (**som**)	3,392	566	1,696	5,654
Sundanese (**sun**)	924	199	926	2,049
Swahili (**swa**)	3,307	551	1,656	5,514
Swedish (**swe**)	1,187	200	1,188	2,575
Tatar (**tat**)	1,000	200	1,000	2,200
Tigrinya (**tir**)	3,681	614	1,840	6,135
Ukrainian (**ukr**)	2,466	249	2,234	4,949
Emakhuwa (**vmw**)	1,551	258	777	2,586
isiXhosa (**xho**)	–	682	1,594	2,276
Yoruba (**yor**)	2,992	497	1,500	4,989
isiZulu (**zul**)	–	875	2,047	2,922

Table 1: Languages and data split sizes. Datasets with no training splits (-) were only used in Track C (crosslingual) only.

ious platforms, including Reddit (e.g., eng, deu), YouTube (e.g., esp, ind, jav, sun, tir), Twitter (e.g., amh, hau), and Weibo (e.g., chn).

- **Personal narratives, talks, speeches**: Anonymised sentences from personal diary posts. We use these in eng, deu, and ptbr, mainly from subreddits such as IAmI. Similarly, the afr dataset includes sentences from speeches and talks.
- **Literary texts**: The language lead manually translated the novel *La Grande Maison* (The Big House) by the Algerian author Mohammed Dib from French into Algerian Arabic (arq), and post-processed the translation to generate sentences for annotation by native speakers. Note that the translator is bilingual and a native speaker of Algerian Arabic.
- **News data**: Although we prefer emotionally rich social media data from different platforms, when such data is scarce, we annotated news data and headlines in some African languages (e.g., yor, hau, and vmw).
- **Human-written and machine-generated data**: We created a dataset from scratch for Hindi (hin) and Marathi (mar). Annotators were asked to come up with emotive sentences on a given topic (e.g., family). A small portion of the Hindi dataset was automatically translated into Marathi and manually corrected by native speakers to fix translation errors. Finally, we augmented both datasets with a few hundred quality-approved instances generated by ChatGPT. Note that these constitute less than 1% of the total number of data instances.

3.2 Data Annotation

We ask the annotators to select all the emotions that apply to a given text. The set of perceived emotion labels includes: *anger, sadness, fear, disgust, joy, surprise*, and *neutral* (if no emotion is present). The annotators further rate the selected emotion(s) on a four-point intensity scale: 0 (no emotion), 1 (low intensity), 2 (moderate intensity), and 3 (high intensity). We provide the definitions of the categories, annotation guidelines, and more details in Muhammad et al. (2025). We expected some level of disagreement, as emotions are complex, subtle, and perceived differently, even by people within the same culture, especially in the absence of full context. Hence, the final emotion labels were determined based on the emotions and associated intensity values selected by the annotators. Specifically, the given emotion is considered present if:

1. At least two annotators select a label with an intensity value of 1, 2, or 3 (low, medium, or high, respectively).
2. The average score exceeds a predefined threshold T. We set T to 0.5.

Once the perceived emotion labels are assigned, the final intensity scores for Track B are determined by averaging the selected intensity values and rounding up to the nearest whole number. Intensity scores are assigned only for datasets in which most instances were annotated by at least five annotators to ensure robustness. Table 1 shows the total number of instances in each dataset, as well as the number of instances in the training, development, and test splits for all languages.

3.3 Annotators' Reliability

We report the reliability of the annotation using the Split-Half Class Match Percentage (SHCMP; Mohammad, 2024) as described in Muhammad et al. (2025). SHCMP extends the concept of Split-Half Reliability (SHR), traditionally used for continuous scores (Kiritchenko and Mohammad, 2016), to discrete categories like ours (i.e., intensity scores per emotion). Overall, the scores vary from 60% to more than 90%, indicating that our datasets are of high quality.

4 Task Description

Participants were given text snippets and asked to determine the emotions that people may attribute to the speaker based on a sentence or short text snippet uttered by the speaker. The task consists of three tracks, and participants could participate in one or more of these tracks.

4.1 Tracks

Track A: Multi-label Emotion Detection Participants were asked to predict the perceived emotion(s) of the speaker and label each text snippet based on the presence (1) or absence (0) of the following emotions: joy, sadness, fear, anger, surprise, and disgust.

Track B: Emotion Intensity Detection Given a text and six emotion classes (i.e., joy, sadness, fear, anger, surprise, and disgust), participants were required to predict whether the intensity of each emotion was 0 (no emotion), 1 (low), 2 (medium), or 3 (high). Note that Track B does not include all languages, as intensity scores were only released for datasets with at least five annotators per instance to ensure more robust and reliable labels.

Track C: Cross-lingual Emotion Detection Similar to Track A, participants were required to predict the presence or absence of each perceived emotion, but without using any training data in the target language. Instead, they were permitted to use labeled dataset(s) from at least one other language. For instance, one could use German data for training when testing on English. This track focuses on cross-lingual transfer and explores how data from various languages can support emotion detection in low-resource settings, as well as the ability of models to generalise across domains.

4.2 Task Organisation

We used Codabench as the competition platform and released pilot datasets before the start of the shared task to help participants better understand the task (i.e., the datasets, the languages involved, and the labels). We provided participants with a starter kit on GitHub, resources for beginners, and organised a Q&A session along with a writing tutorial for junior researchers. Our participants were based in different parts of the world, as shown in Figure 2, with many coming from underrepresented regions. The task consisted of two phases: (1) the development phase and (2) the evaluation phase. During the development phase, the leaderboard was open, allowing a maximum of 999 submissions per participant. In the evaluation phase, the leaderboard was closed, and each participant was allowed up to three submissions, with the last submission being considered for the official ranking.

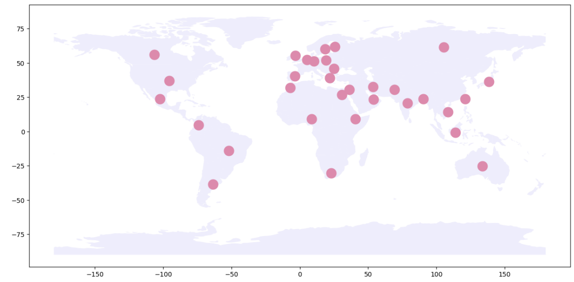

Figure 2: The official affiliations of some of our participants. The list includes 33 countries: Argentina, Australia, Bangladesh, Brazil, Canada, Colombia, Egypt, Ethiopia, Finland, Germany, Greece, India, Indonesia, Iran, Japan, Jordan, Mexico, Morocco, the Netherlands, Nigeria, Pakistan, Poland, Romania, Russia, South Africa, Spain, Sweden, Taiwan, the UAE, the UK, the USA, and Vietnam.

4.3 Evaluation Metrics and Baselines

Evaluation Metrics For Tracks A and C, we use the average macro F-score calculated based on the predicted and the gold-standard labels. For Track B, we use the Pearson correlation coefficient, which captures how well the system-predicted intensity scores of test instances align with human judgments. We provided the participants with an evaluation script on our GitHub page.

Our Baselines We run a simple majority class baseline for each language across all three tracks. Further, for Tracks A and B (Tables 2 and 3, respectively), we fine-tuned RoBERTa using the training data for each language. Table 2 shows the average macro F-scores of the top-performing systems compared to our baseline in Track A, and Table 3 shows the Pearson correlation scores for Track B. For Track C (Table 4), we fine-tuned RoBERTa by training on all languages within a language family while holding out one target language used for testing, e.g., all Indo-European languages except `eng` when testing on it. For language families with only one language, we trained on the Slavic languages (`rus` and `ukr`) and tested on `tat`; on the

Lang	Team	Score	Lang	Team	Score	Lang	Team	Score	Lang	Team	Score
afr	**pai**	**0.699**	amh	**chinchunmei**	**0.773**	arq	**pai**	**0.669**	ary	**pai**	**0.629**
	maomao	**0.687**		**nust titans**	**0.714**		**jnlp**	**0.641**		**jnlp**	**0.609**
	R_{baseline}	0.371		R_{baseline}	0.638		R_{baseline}	0.414		R_{baseline}	0.472
	M_{baseline}	0.257		M_{baseline}	0.295		M_{baseline}	0.445		M_{baseline}	0.247
chn	**pai**	**0.709**	deu	**pai**	**0.740**	eng	**pai**	**0.823**	esp	**pai**	**0.849**
	teleai	**0.682**		**heimerdinger**	**0.706**		**nycu-nlp**	**0.823**		**heimerdinger**	**0.838**
	R_{baseline}	0.531		R_{baseline}	0.642		R_{baseline}	0.708		R_{baseline}	0.774
	M_{baseline}	0.278		M_{baseline}	0.449		M_{baseline}	0.367		M_{baseline}	0.312
hau	**pai**	**0.751**	hin	**jnlp**	**0.926**	ibo	**pai**	**0.600**	kin	**pai**	**0.657**
	empaths	**0.695**		**pai**	**0.920**		**late-gil-nlp**	**0.563**		**mcgill-nlp**	**0.590**
	R_{baseline}	0.596		R_{baseline}	0.855		R_{baseline}	0.479		R_{baseline}	0.463
	M_{baseline}	0.312		M_{baseline}	0.246		M_{baseline}	0.236		M_{baseline}	0.218
mar	**pai**	**0.884**	orm	**tewodros**	**0.616**	pcm	**pai**	**0.674**	ptbr	**pai**	**0.683**
	indidataminer	**0.883**		**late-gil-nlp**	**0.592**		**jnlp**	**0.634**		**heimerdinger**	**0.625**
	R_{baseline}	0.822		R_{baseline}	0.126		R_{baseline}	0.555		R_{baseline}	0.426
	M_{baseline}	0.264		M_{baseline}	0.232		M_{baseline}	0.357		M_{baseline}	0.243
ptmz	**pai**	**0.548**	ron	**pai**	**0.794**	rus	**heimerdinger**	**0.901**	som	**pai**	**0.577**
	heimerdinger	**0.507**		**jnlp**	**0.779**		**jnlp**	**0.891**		**empaths**	**0.508**
	R_{baseline}	0.459		R_{baseline}	0.762		R_{baseline}	0.838		R_{baseline}	0.459
	M_{baseline}	0.163		M_{baseline}	0.461		M_{baseline}	0.262		M_{baseline}	0.198
sun	**lazarus nlp**	**0.550**	swa	**empaths**	**0.386**	swe	**pai**	**0.626**	tat	**pai**	**0.846**
	pai	**0.541**		**pai**	**0.385**		**jnlp**	**0.619**		**tue-jms**	**0.797**
	R_{baseline}	0.373		R_{baseline}	0.227		R_{baseline}	0.520		R_{baseline}	0.539
	M_{baseline}	0.334		M_{baseline}	0.179		M_{baseline}	0.264		M_{baseline}	0.246
tir	**nta**	**0.591**	ukr	**pai**	**0.726**	vmw	**team unibuc**	**0.325**	yor	**pai**	**0.461**
	late-gil-nlp	**0.587**		**csiro-lt**	**0.664**		**pai**	**0.255**		**heimerdinger**	**0.392**
	R_{baseline}	0.463		R_{baseline}	0.535		R_{baseline}	0.121		R_{baseline}	0.092
	M_{baseline}	0.253		M_{baseline}	0.157		M_{baseline}	0.163		M_{baseline}	0.165

Table 2: Average macro-F1 scores for our baselines ($M_{baseline}$ and $R_{baseline}$, referring to the Majority Vote and RoBERTa baselines, respectively) and the top two performing systems in Track A (shown in bold) for each language.

Niger-Congo languages (swa and yor) and tested on pcm; and trained on rus when testing on chn.

5 Participating Systems and Results

5.1 Overview

Our task attracted more than 700 registered participants and was featured in the Codabench newsletter as the most popular competition hosted on Codabench in 2024.

In the development phase, 153 submissions were made for Track A, 52 for Track B, and 25 for Track C. In the test phase, 220 submissions were made for Track A, 96 for Track B, and 46 for Track C. The official results include more than 220 final submissions from 93 teams. While the English subtracks received the highest number of submissions, we note that other languages, including underserved ones, were comparable in terms of popularity.

We report results only for teams that submitted a system description paper. **??** presents the results for Track A, which had 87 participating teams. **??** shows the results for Track B, with 38 participating teams, while **??** reports the results for Track C, which had 21 participating teams.

5.2 Track A: Multi-label Emotion Detection

5.2.1 Best-Performing Systems

Team Pai proposes one of the most effective models in the competition. They consistently rank as the top approach in Track A for 20 out of 28 languages. For their system, they combine several base models (ChatGPT-4o (OpenAI, 2024), DeepSeek-V3 (DeepSeek-AI et al., 2025), Gemma-9b (Team et al., 2024), Qwen-2.5-32b (Yang et al., 2024), Mistral-Small-24B (Jiang et al., 2024)) using multiple ensemble techniques (neural networks, XGBoost, LightGBM, linear regression, weighted voting). They fine-tune Gemma-9b and Qwen-2.5-32b using AdaLoRA. For prompting the LLMs, they used an iterative prompt-optimisation technique that generates prompt variations.

Team Chinchunmei ranks in the top 10 in 16 languages in Track A and 12th in English. They use sample contrastive learning, where performance is enhanced by comparing sample pairs, and generative contrastive learning, where the models learn to distinguish correct from incorrect predictions. Their samples are randomly selected from the task dataset (no external augmentation). They use LLaMa3-Instruct-8B (AI@Meta, 2024) for their fine-tuning.

5.2.2 Takeaways

Most of the teams that rank well on Track A experiment with essentially two methodologies: 1) fine-tuning BERT-based models such as DeBERTa (Siino, 2024), mBERT (Dolev, 2023), and XLM-R (Conneau et al., 2020); and/or 2) instruction-fine-tuning using LoRa methodologies in combination with prompt design and data augmentation techniques on LLMs (ChatGPT-4o, DeepSeek-V3, Gemma-9b, Qwen-2.5-32b, Mistral-Small-24B). For instance, Team Telai (3rd in Chinese), Team Empaths (2nd in Hausa and Somali, and 1st in Swahili), Team NYCU-NLP (2nd in English), Team JNLP (1st in Hindi and among the best four across 10 languages), Team Unibouc (1st in Emakhuwa), Team Heiderdinger (2nd in Mozambican Portuguese, German, Spanish, Brazilian Portuguese, and Yorùbá; 1st in Russian), and Team Maomao (2nd in Afrikaans).

Few teams focus on only a subset of languages or explore language-related knowledge in their methodology. For instance, Team Lazarus NLP redefined and reformulated the multi-label classification into multiple binary tasks to expand training samples. They also explored how knowledge could potentially be transferred between Indonesian languages. All the top 10 teams performed significantly better than our baseline model, with an improvement that is more notable in a few low-resource languages, such as Oromo, where Team Tewodros obtained an average macro-F1 score of 0.616 compared to a baseline of 0.126. The same was observed for Yorùbá, where Team Pai scored 0.461 compared to a baseline score of 0.092.

5.3 Track B: Emotion Intensity Detection

5.3.1 Best-Performing Systems

Team Pai Similar to Track A, Team Pai ranked at the top across all languages in Track B, except for Amharic. They used an ensemble of LLMs, combining several base models (ChatGPT-4o(OpenAI, 2024), DeepSeek-V3 (DeepSeek-AI et al., 2025), Gemma-9b (Team et al., 2024), Qwen-2.5-32b (Yang et al., 2024), Mistral-Small-24B) with multiple ensemble techniques (neural networks, XGBoost, LightGBM, linear regression, weighted voting). They fine-tuned the Gemma and Qwen models using AdaLoRA. For prompting the LLMs, they employed an iterative prompt-optimisation technique to generate prompt variations.

Lang	Team	Score	Lang	Team	Score
amh	**csecu-learners**	**0.856**	arq	**pai**	**0.650**
	heimerdinger	**0.781**		**jnlp**	**0.587**
	R_{baseline}	0.508		R_{baseline}	0.016
	M_{baseline}	-0.001		M_{baseline}	-0.009
chn	**pai**	**0.722**	deu	**pai**	**0.766**
	teleai	**0.708**		**teleai**	**0.743**
	R_{baseline}	0.405		R_{baseline}	0.562
	M_{baseline}	0.000		M_{baseline}	0.016
eng	**pai**	**0.840**	esp	**pai**	**0.808**
	nycu-nlp	**0.837**		**deepwave**	**0.792**
	R_{baseline}	0.641		R_{baseline}	0.726
	M_{baseline}	0.001		M_{baseline}	0.011
hau	**pai**	**0.770**	ptbr	**pai**	**0.710**
	deepwave	**0.747**		**teleai**	**0.690**
	R_{baseline}	0.270		R_{baseline}	0.297
	M_{baseline}	0.003		M_{baseline}	0.016
ron	**pai**	**0.726**	rus	**pai**	**0.925**
	deepwave	**0.716**		**teleai**	**0.919**
	R_{baseline}	0.557		R_{baseline}	0.877
	M_{baseline}	0.003		M_{baseline}	0.016
ukr	**pai**	**0.708**			
	jnlp	**0.672**			
	R_{baseline}	0.399			
	M_{baseline}	-0.01			

Table 3: Pearson correlation scores for our baselines (Majority: $M_{baseline}$ and RoBERTa: $R_{baseline}$) and the top two performing systems in Track B (shown in bold) for each language.

Team CSECU-Learners CSECU-Learners ranked at the top in Amharic by fine-tuning language-specific transformers (XLM-Roberta (Conneau et al., 2020) for Amharic) with a classification layer and multi-sample dropout.

5.3.2 Takeaways

Teams Deepwave, Teleai, and JNLP also ranked highly across various languages using prompt engineering approaches similar to those in Track A. Additionally, Team NYCU-NLP ranked second in English by aggregating instruction-tuned small language models. All these teams outperformed our RoBERTa baseline, which achieved moderate Pearson correlation coefficient scores overall, but performed poorly in languages such as Algerian Arabic, Hausa, Ukrainian, and even Brazilian Portuguese -highlighting the difficulty of the task.

Overall, we observe that most teams adopted approaches similar to those used in Track A, with only minor adjustments to the prompts. Notably, even the best-performing teams achieved a Pearson correlation coefficient of no more than 0.65

Lang	Team	Score	Lang	Team	Score	Lang	Team	Score	Lang	Team	Score
afr	maomao	**0.705**	amh	**deepwave**	**0.661**	arq	**deepwave**	**0.588**	ary	**deepwave**	**0.632**
	deepwave	**0.574**		**uob-nlp**	**0.627**		**maomao**	**0.584**		**maomao**	**0.565**
	R_{baseline}	0.350		R_{baseline}	0.487		R_{baseline}	0.338		R_{baseline}	0.355
	M_{baseline}	0.257		M_{baseline}	0.295		M_{baseline}	0.445		M_{baseline}	0.247
chn	**deepwave**	**0.689**	deu	**deepwave**	**0.727**	eng	**deepwave**	**0.797**	esp	**deepwave**	**0.831**
	maomao	**0.622**		**gt-nlp**	**0.687**		**maomao**	**0.755**		**maomao**	**0.806**
	R_{baseline}	0.246		R_{baseline}	0.468		R_{baseline}	0.375		R_{baseline}	0.574
	M_{baseline}	0.278		M_{baseline}	0.319		M_{baseline}	0.449		M_{baseline}	0.367
hau	**deepwave**	**0.709**	hin	**deepwave**	**0.919**	ibo	**deepwave**	**0.605**	ind	**maomao**	**0.672**
	uob-nlp	**0.627**		**maomao**	**0.896**		**uob-nlp**	**0.484**		**lazarus nlp**	**0.641**
	R_{baseline}	0.320		R_{baseline}	0.138		R_{baseline}	0.075		R_{baseline}	0.376
	M_{baseline}	0.312		M_{baseline}	0.264		M_{baseline}	0.236		M_{baseline}	0.254
jav	**heimerdinger**	**0.439**	kin	**deepwave**	**0.508**	mar	**deepwave**	**0.903**	orm	**deepwave**	**0.542**
	lazarus nlp	**0.438**		**uob-nlp**	**0.466**		**maomao**	**0.863**		**uob-nlp**	**0.491**
	R_{baseline}	0.464		R_{baseline}	0.184		R_{baseline}	0.772		R_{baseline}	0.262
	M_{baseline}	0.204		M_{baseline}	0.218		M_{baseline}	0.264		M_{baseline}	0.232
pcm	**deepwave**	**0.674**	ptbr	**deepwave**	**0.629**	ptmz	**deepwave**	**0.555**	ron	**deepwave**	**0.767**
	maomao	**0.562**		**maomao**	**0.617**		**maomao**	**0.495**		**maomao**	**0.747**
	R_{baseline}	0.010		R_{baseline}	0.418		R_{baseline}	0.297		R_{baseline}	0.762
	M_{baseline}	0.357		M_{baseline}	0.243		M_{baseline}	0.163		M_{baseline}	0.652
rus	**deepwave**	**0.906**	som	**maomao**	**0.488**	sun	**deepwave**	**0.467**	swa	**maomao**	**0.381**
	maomao	**0.852**		**deepwave**	**0.488**		**maomao**	**0.464**		**deepwave**	**0.355**
	R_{baseline}	0.704		R_{baseline}	0.273		R_{baseline}	0.194		R_{baseline}	0.190
	M_{baseline}	0.262		M_{baseline}	0.198		M_{baseline}	0.334		M_{baseline}	0.179
swe	**deepwave**	**0.645**	tat	**deepwave**	**0.789**	tir	**deepwave**	**0.505**	ukr	**deepwave**	**0.702**
	maomao	**0.578**		**maomao**	**0.697**		**uob-nlp**	**0.445**		**maomao**	**0.623**
	R_{baseline}	0.512		R_{baseline}	0.445		R_{baseline}	0.339		R_{baseline}	0.496
	M_{baseline}	0.264		M_{baseline}	0.246		M_{baseline}	0.253		M_{baseline}	0.157
vmw	**deepwave**	**0.210**	xho	**maomao**	**0.443**	yor	**maomao**	**0.360**	zul	**maomao**	**0.397**
	ozemi	**0.193**		**ozemi**	**0.315**		**deepwave**	**0.342**		**heimerdinger**	**0.226**
	R_{baseline}	0.052		R_{baseline}	0.127		R_{baseline}	0.053		R_{baseline}	0.153
	M_{baseline}	0.162		M_{baseline}	0.115		M_{baseline}	0.165		M_{baseline}	0.109

Table 4: Average macro-F1 scores for our baselines ($M_{baseline}$ and $R_{baseline}$, referring to the Majority Vote and RoBERTa baselines, respectively) and the top two performing systems in Track C (shown in bold) for each language.

on Algerian Arabic, likely due to the novelty and complexity of the dataset.

5.4 Track C: Cross-lingual Emotion Detection

5.4.1 Best-Performing Systems

Team deepwave Team Deepwave fine-tuned Google Gemma-2 (Team et al., 2024) using tailored data augmentation and Chain-of-Thought (CoT) prompting. They decomposed the task into two sub-tasks: (1) sentiment keyword identification and (2) sentiment polarity recognition.

To address the challenge of limited data, they employed k-fold (k=5) cross-validation and used model merging—a strategy that combines the predictions of multiple models to improve generalization—by averaging the prediction probabilities of each model, assigning equal weights of 0.2. In this track, a dedicated LoRA module was trained for each target language. The training dataset for each module comprised data from all other languages in Track A, excluding the target language L_i. They exclusively used augmented data generated through CoT prompting for training.

Team maomao Team maomao experimented with different setups for fine-tuning LLMs. The base models used were Qwen2.5-7B-Instruct, GPT-0544o Mini2, and LLaMA-3.2-3B-Instruct (AI@Meta, 2024). They applied Direct Preference Optimisation to refine their model -a technique that selects high-quality instances from lower-quality ones within a dataset. After this step, they retrained the model using the refined dataset. They also explored random sampling and retrieval-augmented generation (RAG) methods for training, primarily on DeepSeek-V3, Qwen-Max5093, and Grok-V2.

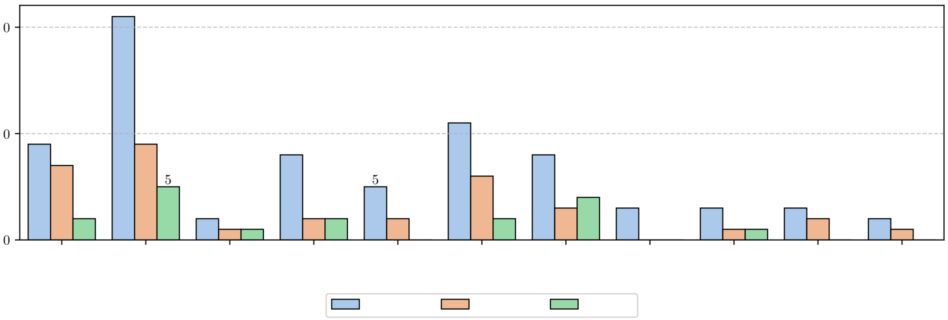

Figure 3: Top LLMs used by participants across tracks (A, B, and C).

5.4.2 Takeaways

Other top-performing systems include Team Ozemi, who fine-tuned a multilingual BERT model and applied machine translation to enhance performance across all languages. They used the Synthetic Minority Oversampling Technique (SMOTE) with TF-IDF to address class imbalance in Russian. They translated all the datasets into a common language using Google Translate before processing -except for Nigerian Pidgin and Emakhua, where they used a multilingual BERT model for translation. They also leveraged two Kaggle competition datasets for data augmentation.

Team heimerdinger, who ranked highest in Javanese, built their approach using various LLMs (LLaMA 3.1 8B, Qwen 2.5 7B, DeepSeek-7B, MistralV0.3-7B, and Gemma2-9B) for Track C. They employed in-context learning with multilingual examples from high-resource languages such as English and Spanish.

Overall, the participating teams outperformed our baselines. However, the average scores for this track are notably lower, particularly in low-resource languages, due to the additional challenges posed by limited data and resources. As shown in Table 4, there are significant performance gaps -even the top systems did not achieve an F-score higher than 0.50 in languages such as Javanese, Somali, Sundanese, Xhosa, Yorùbá, isiZulu, and Emakhuwa (where the top system achieved an F-score of only 0.21).

6 Discussion

Popular Methods Unsurprisingly, most top-performing teams favored fine-tuning and prompting large language models (LLMs) such as Gemma-2, Mistral, Phi-4, Qwen-2.5, DeepSeek, LLaMA-3, GPT, and Gemini models. For fine-tuning, both full fine-tuning and parameter-efficient fine-tuning were the most commonly used strategies to enhance performance.

For prompting, few-shot, zero-shot, and chain-of-thought prompting were the most frequently used techniques.

Many participants also experimented with traditional transformer-based models, particularly XLM-RoBERTa, mBERT, DeBERTa, and IndicBERT (Kakwani et al., 2020) (see Figure 3 and **??** in the Appendix).

Best Performing Systems The results from the top-performing submissions suggest that while LLMs achieve strong overall performance, their effectiveness is heavily dependent on prompt engineering techniques and wording.

Additionally, performance varies significantly by language. Across all tracks, LLM-based approaches and the best-performing systems consistently yielded better results for high-resource languages such as English and Russian. In contrast, performance dropped notably when tested on low-resource languages such as Swahili and Emakhuwa.

Furthermore, most teams did not incorporate additional datasets to enhance performance (see Appendix), as few-shot and zero-shot approaches proved highly effective.

7 Conclusion

We presented our shared task on text-based emotion recognition, which covered three tracks and a total of 32 languages. The submitted systems were ranked based on macro F1-scores for Tracks A and C, comparing predicted labels to gold labels, and based on the ranking of predicted intensity scores for Track B.

We summarised the reported results, discussing the best-performing and most innovative methods.

Overall, performance varied significantly across languages. Our results highlight that emotion recognition remains an open challenge, particularly for under-served languages and in low-resource settings.

8 Limitations

Emotions are subjective and subtle, and they are expressed and perceived differently. We do not claim that our datasets capture the true emotions of the speakers, fully represent language use across the 32 languages, or cover all possible emotions. We discuss the ethical considerations extensively in Section 9.

We acknowledge the limited data sources available for some low-resource languages. Therefore, our datasets may not be suitable for tasks requiring large amounts of data in a given language. However, they serve as a valuable starting point for research in this area.

9 Ethical Considerations

Emotion perception and expression are subjective and nuanced, as they are influenced by various factors (e.g., cultural background, social group, personal experiences, and social context). Thus, it is impossible to determine someone's emotions with absolute certainty based solely on short text snippets. Our datasets explicitly focus on perceived emotions—identifying the emotions that most people believe the speaker may have felt. We do not claim to annotate the speaker's true emotions, as these cannot be definitively determined from text alone. We recognise the importance of this distinction, as perceived emotions may differ from actual emotions.

We acknowledge potential biases in our data, as we rely on text-based communication, which inherently carries biases from data sources and annotators. Additionally, while many of our datasets focus on low-resource languages, we do not claim they fully represent these languages' usage. Further, although we took measures to filter inappropriate content, some instances may have been overlooked.

We explicitly urge careful consideration of ethical implications before using our datasets. We prohibit their use for commercial purposes or by state actors in high-risk applications unless explicitly approved by the dataset creators. Systems built on our datasets may not be reliable at the individual instance level and are susceptible to domain shifts. Thus, they should not be used for critical decision-making, such as in health applications, without expert supervision. For a more in-depth discussion, see Mohammad (2022, 2023).

Finally, all annotators involved in the study were compensated above the minimum hourly wage.

Acknowledgments

Shamsuddeen Muhammad acknowledges the support of Google DeepMind and Lacuna Fund, an initiative co-founded by The Rockefeller Foundation, Google.org, and Canada's International Development Research Centre. The views expressed herein do not necessarily represent those of Lacuna Fund, its Steering Committee, its funders, or Meridian Institute.

Nedjma Ousidhoum would like to thank Abderrahmane Samir Lazouni, Lyes Taher Khalfi, Manel Amarouche, Narimane Zahra Boumezrag, Noufel Bouslama, Abderaouf Ousidhoum, Sarah Arab, Wassil Adel Merzouk, Yanis Rabai, and another annotator who would like to stay anonymous for their work and insightful comments.

Idris Abdulmumin gratefully acknowledges the ABSA UP Chair of Data Science for funding his post-doctoral research and providing compute resources.

Jan Philip Wahle and Terry Ruas were partially supported by the Lower Saxony Ministry of Science and Culture and the VW Foundation.

Meriem Beloucif acknowledges Nationella Språkbanken (the Swedish National Language Bank) and Swe-CLARIN – jointly funded by the Swedish Research Council (2018–2024; dnr 2017-00626) and its 10 partner institutions for funding the Swedish annotations.

Alexander Panchenko would like to thank Nikolay Ivanov, Artem Vazhentsev, Mikhail Salnikov, Maria Marina, Vitaliy Protasov, Sergey Pletenev, Daniil Moskovskiy, Vasiliy Konovalov, Elisey Rykov, and Dmitry Iarosh for their help with the annotation for Russian. Preparation of Tatar data was funded by AIRI and completed by Dina Abdullina, Marat Shaehov, and Ilseyar Alimova.

Yi Zhou would like to thank Gaifan Zhang, Bing Xiao, and Rui Qin for their help with the annotations and for providing feedback.

References

AI@Meta. 2024. Llama 3 model card.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2021. FEEL-IT: Emotion and sentiment classification for the Italian language. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 76–83, Online. Association for Computational Linguistics.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2022. Xlm-emo: Multilingual emotion prediction in social media text. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 195–203.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alexandra Ciobotaru, Mihai Vlad Constantinescu, Liviu P. Dinu, and Stefan Dumitrescu. 2022. RED v2: Enhancing RED dataset for multi-label emotion detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1392–1399, Marseille, France. European Language Resources Association.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Eyal Liron Dolev. 2023. Does mBERT understand Romansh? evaluating word embeddings using word alignment. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 41–53, Neuchatel, Switzerland. Association for Computational Linguistics.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

MD Asif Iqbal, Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H Sarker. 2022. Bemoc: A corpus for identifying emotion in bengali texts. *SN Computer Science*, 3(2):135.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego, California.

Youssef Mohamed, Runjia Li, Ibrahim Said Ahmad, Kilichbek Haydarov, Philip Torr, Kenneth Church, and Mohamed Elhoseiny. 2024. No culture left behind: ArtELingo-28, a benchmark of WikiArt with captions in 28 languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20939–20962, Miami, Florida, USA. Association for Computational Linguistics.

Saif Mohammad. 2022. Ethics sheets for AI tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8368–8379, Dublin, Ireland. Association for Computational Linguistics.

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 Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.

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

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018b. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. 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. 2024. WorryWords: Norms of anxiety association for over 44k English words. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16261–16278, Miami, Florida, USA. Association for Computational Linguistics.

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. 2025. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv preprint arXiv:2502.11926*.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. Xed: A multilingual dataset for sentiment analysis and emotion detection. *arXiv preprint arXiv:2011.01612*.

OpenAI. 2024. Gpt-4o system card.

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

Mei Silviana Saputri, Rahmad Mahendra, and Mirna Adriani. 2018. Emotion classification on indonesian twitter dataset. In *2018 International Conference on Asian Language Processing (IALP)*, pages 90–95.

Marco Siino. 2024. DeBERTa at SemEval-2024 task 9: Using DeBERTa for defying common sense. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 291–297, Mexico City, Mexico. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.