Advancing Emotion Recognition and Intensity Modeling for Ethiopian Languages

Tadesse Destaw Belay

Instituto Politécnico Nacional, Mexico

Contact: tadesseit@gmail.com

Abstract

Understanding emotions is fundamental to many human-computer interaction tasks, including customer feedback analysis, marketing insights, and social media monitoring. In real-world scenarios, individuals often express multiple emotions simultaneously, making multi-label annotation essential for capturing emotional complexity. While the **EthioEmo** dataset (Belay et al., 2025) provides multi-label emotion annotations for Ethiopian languages, it lacks information on emotion intensities, which are crucial for modeling the strength or subtlety of each emotional expression. In this work, we extend the EthioEmo dataset by introducing emotion intensity annotations for each labeled emotion, offering a richer and more nuanced resource for emotion understanding in low-resource African languages. We evaluate a range of encoder-only Pretrained Language Models (PLMs) and open-source Large Language Models (LLMs) on this enhanced dataset. Preliminary results show that African-language PLMs consistently outperform open-source LLMs, underscoring the importance of culturally and linguistically aligned models for emotion analysis in underrepresented languages.

1 Introduction

Emotion detection is one of the most challenging and subjective tasks in Natural Language Processing (NLP) (Ziems et al., 2024). Unlike many other NLP tasks, it requires assigning a text to the emotion label(s) that most accurately reflect the mental state of the author or a reader. The ability to detect emotions in text has numerous applications, from identifying (dis)satisfaction in customer feedback to evaluating the emotional well-being of individuals and societies (Liu, 2012).

There are two approaches to annotate an emotion dataset: *multi-class* and *multi-label*. In the *multi-class* approach, a text is assigned to either a single emotion class or no emotion. In contrast, the *multi-label* approach allows a text to be associated with none, one, multiple, or all of the targeted emotion labels. Furthermore, emotion intensity is an extension of the emotion detection task that quantifies the strength of each expressed emotion (Mashal and Asnani, 2017). In multi-label emotion, adding the intensity of each corresponding selected emotion is crucial, as each emotion might not always be equally expressed (Labat et al., 2022; Firdaus et al., 2020). For instance, some feelings may be subtly present, while others dominate more prominently. This complexity highlights the importance of assessing intensity, as it provides a nuanced understanding of how emotions are expressed. Consider the sentence, *'Although I'm incredibly excited about starting my new job, I feel a little sad about leaving my friends I made there.'* Here, the sense of happiness (joy) is pronounced and primary, whereas the feeling of sadness is secondary and less intense. As illustrated in Figure 1, some texts have a single emotion label with its corresponding intensity value, while others have multiple emotions, each with its intensity levels.

The work by Belay et al. (2025) created EthioEmo, a multi-label emotion dataset for four Ethiopian low-resource languages, namely Amharic (`amh`), Oromo (`orm`), Somali (`som`), and Tigrinya (`tir`). However, this multi-label emotion dataset is annotated without considering the intensity of the labeled emotions. This work contributes by: 1) extending the EthioEmo dataset to incorporate intensity of the labeled emotions, thereby enriching the applicability of the dataset for nuanced emotion analysis, and 2) evaluating BERT-based pre-trained language models (PLMs) along with open-source large language models (LLMs) for their effectiveness in multi-label emotion classification, intensity prediction, and exploring the feasibility of cross-lingual transfer learning across the four Ethiopian languages.

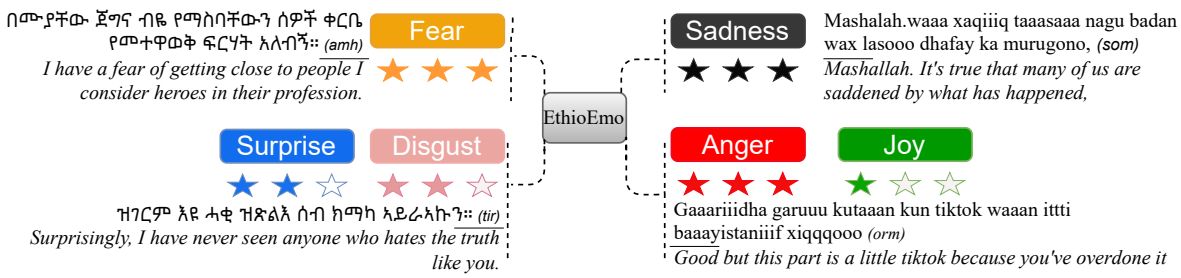

Figure 1: Multi-label emotion in Ethiopian languages (EthioEmo)dataset with its level of intensity. The text might have one, two, many, or all emotions with its level of intensity for the selected emotion.

2 Related Work

Multi-label Emotion: Emotion is central to human nature, and as online interactions grow, people express and react to a content in various ways. A text expression can simultaneously manifest multiple emotions to reflect the complex emotional nuances conveyed (Mashal and Asnani, 2017). To handle the complex multiple emotion expression, some of the recent multi-label emotion datasets are SemEval-2018 Task 1 (Mohammad et al., 2018), GoEmotions (Demszky et al., 2020), EmoInHindi (Singh et al., 2022), WASSA-2024 shared task dataset (Giorgi et al., 2024), BRIGHTER (Muhammad et al., 2025a), EthioEmo (Belay et al., 2025), and SemEval2025 Task 11 data (Muhammad et al., 2025b).

Intensity in Multi-label Emotion: In addition to emotion classification, analyzing the degree of each emotion provides deeper insights, leading to more informed and effective decisions (Maruf et al., 2024). Accurately annotating the intensity of each labeled emotion is essential for advancing the capabilities of language models, as it presents an additional challenge for nuanced emotion recognition. Most common multi-label emotion datasets (Mohammad et al., 2018; Singh et al., 2022; Giorgi et al., 2024; Muhammad et al., 2025a) include intensity ratings for the corresponding emotion. However, the EthioEmo (Belay et al., 2025) dataset is annotated in a multi-label annotation setting without specifying the intensity of each corresponding emotion. Inspired by this work, we extend the EthioEmo dataset by adding an intensity feature.

Cross-Lingual Experimentation: Cross-lingual transfer learning has emerged as a promising approach to overcome the data scarcity issue in low-resource languages (Maladry et al., 2024). It has been used to transfer knowledge from high-resource to low-resource languages (Zhang et al., 2024). By utilizing cross-lingual approaches, one language can benefit from the resources and insights of another, thus enhancing model generalization over emotion-related tasks (Zhu et al., 2024; Kadiyala, 2024; Cheng et al., 2024). Navas Alejo et al. (2020) explored various cross-lingual strategies for emotion detection and intensity grading, illustrating how models can adapt across different languages. However, the evaluation of cross-lingual transfer across different languages spoken within the same country has not been extensively studied. In this work, we conduct cross-lingual evaluations among Ethiopian languages.

3 EthioEmo Dataset

The EthioEmo emotion dataset is an emotion dataset that covers four Ethiopian languages (Belay et al., 2025). Each instance of the dataset is annotated by three annotators, except for Amharic (amh), which is annotated by five. The final label is determined through a majority vote. Regarding the writing script, amh and tir use Ethiopic (Ge'ez) script, and som and orm use Latin. The distribution of the emotions and intensities in the dataset is presented in Appendix A.5 and A.6, respectively.

Emotion Intensity The degree of the feeling in the emotion dataset is vital for the accurate understanding of complex emotions (Firdaus et al., 2020). We enhance the EthioEmo dataset by including annotations for the intensity of each identified emotion. Annotators are trained to assign an intensity label to each emotion category identified. We follow the emotion intensity scaling approaches from previous works (Mohammad et al., 2018; Singh et al., 2022; Muhammad et al., 2025b) and the intensity scale comprises four levels: 0 (No intensity for any of the emotion classes), 1 (Slight), 2 (Mod-

erate), and 3 (High). The final intensity score for each emotion is aggregated and determined using the following formula, with at least two annotators select an intensity value of 1, 2, or 3. The agreement between annotators is in Appendix A.1.

For five annotators (anno) per instance such as amh, the intensity (I) of each emotion is decided by:

$$I_{\text{final}} = \begin{cases} 0 & \text{if Avg} \leq 1.5 \text{ and anno} \geq 2 \\ 1 & \text{if Avg} \in [0.6, 1.5) \text{ and anno} \geq 2 \\ 2 & \text{if Avg} \in [1.5, 2.5) \text{ and anno} \geq 2 \\ 3 & \text{if Avg} \geq 2.5 \text{ and anno} \geq 2 \end{cases}$$

For three annotators per instance:

$$I_{\text{final}} = \begin{cases} 0, & \text{if } 0 \leq \text{Avg} < 1 \\ 1, & \text{if } 1 \leq \text{Avg} \leq 1.5 \\ 2, & \text{if } 1.5 < \text{Avg} \leq 2.5 \\ 3, & \text{if Avg} \geq 2.5 \end{cases}$$

4 Experiment Setup

We select language models for evaluation from different perspectives, such as general multilingual, Africa-centric, and open-source LLMs.

General Multilingual PLMs We evaluate the most common multilingual BERT-like PLMs such as LaBSE (Feng et al., 2022), RemBERT (Chung et al., 2020), XLM-RoBERTa (Conneau et al., 2020), mBERT (Libovický et al., 2019), and mDeBERTa (He et al., 2021).

Africa-centric PLMs We experiment with fine-tuning the most common African-centric language models such as AfriBERTa (Ogueji et al., 2021), AfroLM (Dossou et al., 2022), AfroXLMR (61 and 76 languages) (Alabi et al., 2022), EthioLLM (Tonja et al., 2024).

Large Language Models (LLMs) Based on their popularity in the open-source community, we evaluate the following open-source LLMs: Qwen2.5-72B (Qwen et al., 2025), Dolly-v2-12B (Conover et al., 2023), Llama-3.3-70B (Grattafiori et al., 2024), Mistral-8x7B (Jiang et al., 2024), and DeepSeek-R1-70B (Guo et al., 2025). Appendix A.7 shows all the model details and versions.

Evaluation Setup We finetune encoder-only models using the train-test split of the dataset for emotion classification, intensity prediction, and cross-lingual transfer experiments. For the LLMs, we prompt the LLMs to perform Chain-of-Thought (CoT) and predict the presence of each emotion from a predefined set in a zero-shot setup. The fine-tuning hyperparameters, Appendix A.2, and prompts of LLMs are presented in Appendix A.7.

5 Experiment Results

5.1 Multi-Label Emotion Classification

Table 1 shows the result of multi-label emotion classification. BERT-like encoder-only models achieve better results than LLMs. EthioLLM, a model designed explicitly for Ethiopian languages, is expected to perform well with fine-tuning. However, its performance does not lead to better results. This largely depends on the parameter size and types of training data. AfroXLMR was trained on a $\approx$ 12GB corpus and includes more languages, which makes an effective cross-lingual transfer and better generalization. In contrast, EthioLLM is trained on a $\approx$3GB of less diverse corpus and language coverage. Evaluated LLMs perform worse for low-resource languages, for which Dolly-v2-12B performs the worst, and Llama-3.3-70B performs better comparatively. AfroXLM-R (61L) achieves state-of-the-art results; for instance, amh achieves 68.5% F1 score, while the benchmark result was 67%. Compared to other Ethiopian languages, Amharic (amh) is better represented among the explored language models.

Models	amh	orm	som	tir	Avg.
Monolingual emotion classification					
LaBSE	66.51	41.49	43.99	48.88	50.22
RemBERT	60.15	47.54	48.31	50.37	51.59
mBERT	26.51	40.32	27.01	25.72	29.89
mDeBERTa	53.43	32.84	36.86	41.73	41.22
XLM-RoBERTa	63.73	37.42	33.51	13.32	37.00
EthioLLM	58.68	47.95	33.84	44.78	46.31
AfriBERTa	60.64	54.10	44.66	47.97	53.34
AfroLM	54.76	42.21	32.77	38.60	42.09
AfroXLM-R-61L	67.93	**51.73**	**49.31**	**54.96**	**55.98**
AfroXLM-R-76L	**68.46**	49.68	49.25	53.08	55.11
Zero-shot emotion prediction from LLMs					
Dolly-v2-12B	5.10	22.89	19.82	1.46	12.32
Mistral-8x7B	29.00	24.25	25.63	27.16	26.51
Qwen2.5-72B	37.82	31.56	28.55	31.13	32.27
DeepSeek-R1-70B	36.89	28.15	26.56	26.49	29.52
Llama-3.3-70B	**42.84**	**29.84**	**32.49**	**32.93**	**34.53**

Table 1: Multi-label emotion classification results (Macro F1 score). BERT-like models are finetuned using the dataset train-test split, and results are the average of 5 runs. LLM's results are from zero-shot evaluations.

Models	Intensity prediction (Pearson r)					Cross-lingual emotion transfer (F1)				
	amh	orm	som	tir	Avg.	amh	orm	som	tir	Avg.
LaBSE	47.79	16.53	25.70	32.10	30.53	44.11	20.77	35.18	40.13	35.55
RemBERT	52.73	24.15	24.85	37.63	34.84	42.65	20.87	31.32	33.39	31.81
mBERT	00.00	17.88	5.51	3.13	6.63	25.10	10.79	14.13	18.27	17.07
mDeBERTa	33.07	7.27	7.02	19.24	16.15	36.40	26.63	18.83	38.03	29.97
XLM-RoBERTa	53.63	17.34	18.39	15.95	26.33	23.52	23.69	26.98	38.63	28.21
EthioLLM	41.90	21.58	9.96	22.77	24.05	38.37	22.46	22.76	33.08	30.42
AfriBERTa	39.38	25.24	20.63	27.56	28.20	46.28	35.86	30.81	38.05	37.75
AfroLM	37.75	15.90	5.08	18.42	19.25	32.12	10.38	9.00	25.48	19.25
AfroXLM-R-61L	55.19	26.75	37.81	**41.96**	40.43	56.41	43.24	**42.21**	52.70	48.64
AfroXLM-R-76L	**60.24**	**29.15**	**41.36**	40.32	**42.77**	**56.65**	**45.01**	41.24	**53.39**	**49.07**

Table 2: Emotion intensity prediction (left) results using Pearson correlation and cross-lingual transfer learning using Macro F1 result (right). The best performance scores are highlighted in **bold**.

5.2 Emotion Intensity Prediction

Table 2, *intensity prediction* column reports the results of the intensity Pearson correlation. As all Ethiopian languages are not included during pretraining, mBERT performs worse; the slightly better performance on `orm` and `som` might be since these languages use the Latin script and share some vocabulary. Likewise, in the emotion classification task, AfroXLM-R (76L) achieves better results in intensity prediction. LLMs at intensity prediction are worse than the emotion classification task and are not presented in the table. For instance, in `amh` intensity prediction, Qwen2.5-72B performs 21.15%, Dolly-v2-12B 4.32%, Llama-3.3-70B 33.93%, Mistral-8x7B 13.22%, and DeepSeek-R1-70B performs 29.08%. The subjectivity and complexity of emotion intensity prediction pose a greater challenge even for high-resource languages (Muhammad et al., 2025a). Moreover, emotion intensity prediction requires assessing the strength or degree of an emotion in the text, making it a more subjective and complex task than emotion classification.

5.3 Cross-lingual Emotion Classification

Our experimental setup for cross-lingual transfer involves fine-tuning BERT-like models on all language datasets except the evaluated target language and testing on the held-out language. Table 2, *Cross-lingual emotion transfer* column reports the results of the cross-lingual transfer learning. AfroXLM-R (76L) achieves better results for the cross-lingual evaluations because it includes all the targeted Ethiopian languages during pretraining. When we compare the cross-lingual results, `amh` and `tir` perform better in transferring. This might be due to both using the same Ethiopic script and having more linguistic corpus used in the pretraining of the base models. From BERT-like models, mBERT performs the worst as none of the languages are included during pretraining. The AfroLM model performs the second worst; it includes only `amh` in the pretraining. In the cross-lingual transfer learning experiments, we observe that excluding Latin script languages during fine-tuning does not impact the performance of Ethiopic script languages. For instance, excluding the Latin script languages, `orm` and `som`, training with the Ethiopic script is effective, showing that they are non-transferable.

6 Conclusion

In this work, we extend the EthioEmo emotion dataset by adding the intensity of the corresponding labeled emotions. Using the dataset, we experiment with multi-label emotion classification, emotion intensity prediction, and cross-lingual transfer learning among Ethiopian languages. Generally, the African-centric language model (AfroXLMR) is better for emotion, intensity, and transferability between Ethiopian languages. This dataset will contribute to developing a more robust emotion evaluation task for low-resource languages. In future work, we suggest modeling the annotator-level data instead of making the majority vote, as making the majority vote does not consider the minority perspectives of annotators when deciding the final gold label for subjective NLP tasks.

Limitations

Limited Annotators per instance While it is common to annotate multi-label emotion using three raters per instance, such as the GoEmotions dataset (Demszky et al., 2020), WRIME emotion intensity (Kajiwara et al., 2021), and others, it is recommended that the more annotators, the higher the quality of the dataset (Troiano et al., 2021; Suzuki et al., 2022). For instance, the BRIGHTER emotion (Muhammad et al., 2025a) dataset intensity of the corresponding emotion is annotated by a minimum of five annotators. Based on our scope, we annotate the intensity using only a minimum of three raters per instance, except amh language, which has five annotators per instance. Future work can add more annotations on top of our three annotators for a better quality of emotion intensity.

Majority vote limitation Regarding deciding the final intensity, we determined the intensity label using majority vote and threshold average of the intensity values. This approach may not incorporate all the perspectives of annotators, as it is the general drawback of the majority vote. We plan to make publicly available the annotator-level data and open it for further exploration of deciding the final emotion intensity, or it can be used to model without applying a majority vote.

Limited model evaluation We evaluated limited LLMs in a zero-shot setup based on our resource limitations. This evaluation can be extended by including more open-source LLMs, proprietary LLMs, and a few-shot evaluation setup.

Ethical Considerations

As we start from a previously annotated dataset (Belay et al., 2025), emotion intensity annotation, perception, and expression are subjective and nuanced as they are strongly related to sociodemographic aspects (e.g., cultural background, social group, personal experiences, social context). Thus, we can never truly identify how one is feeling based solely on the given text snippets with absolute certainty. We ensure fair and honest analysis while conducting our work ethically and without harming anybody.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

Long Cheng, Qihao Shao, Christine Zhao, Sheng Bi, and Gina-Anne Levow. 2024. TEII: Think, explain, interact and iterate with large language models to solve cross-lingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 495–504, Bangkok, Thailand. Association for Computational Linguistics.

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.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic

BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4441–4453, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Salvatore Giorgi, João Sedoc, Valentin Barriere, and Shabnam Tafreshi. 2024. Findings of WASSA 2024 shared task on empathy and personality detection in interactions. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 369–379, Bangkok, Thailand. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

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, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Ram Mohan Rao Kadiyala. 2024. Cross-lingual emotion detection through large language models. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 464–469, Bangkok, Thailand. Association for Computational Linguistics.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104, Online. Association for Computational Linguistics.

Sofie Labat, Naomi Ackaert, Thomas Demeester, and Veronique Hoste. 2022. Variation in the expression and annotation of emotions: A Wizard of Oz pilot study. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 66–72, Marseille, France. European Language Resources Association.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *Preprint*, arXiv:1911.03310.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*, volume 5 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Aaron Maladry, Pranaydeep Singh, and Els Lefever. 2024. Findings of the WASSA 2024 EXALT shared task on explainability for cross-lingual emotion in tweets. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 454–463, Bangkok, Thailand. Association for Computational Linguistics.

Abdullah Al Maruf, Fahima Khanam, Md. Mahmudul Haque, Zakaria Masud Jiyad, M. F. Mridha, and Zeyar Aung. 2024. Challenges and opportunities of text-based emotion detection: A survey. *IEEE Access*, 12:18416–18450.

Sonia Xylina Mashal and Kavita Asnani. 2017. Emotion intensity detection for social media data. In *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, pages 155–158.

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, New Orleans, Louisiana. 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. 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.

Irean Navas Alejo, Toni Badia, and Jeremy Barnes. 2020. Cross-lingual emotion intensity prediction. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 140–152, Barcelona, Spain (Online). Association for Computational Linguistics.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. 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.

Gopendra Vikram Singh, Priyanshu Priya, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. EmoInHindi: A multi-label emotion and intensity annotated dataset in Hindi for emotion recognition in dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5829–5837, Marseille, France. European Language Resources Association.

Haruya Suzuki, Sora Tarumoto, Tomoyuki Kajiwara, Takashi Ninomiya, Yuta Nakashima, and Hajime Nagahara. 2022. Emotional intensity estimation based on writer's personality. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 1–7, Online. Association for Computational Linguistics.

Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Ah Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, Dietrich Klakow, and Seid Muhie Yimam. 2024. EthioLLM: Multilingual large language models for Ethiopian languages with task evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6341–6352, Torino, Italia. ELRA and ICCL.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2021. Emotion ratings: How intensity, annotation confidence and agreements are entangled. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 40–49, Online. Association for Computational Linguistics.

Jinghui Zhang, Yuan Zhao, Siqin Zhang, Ruijing Zhao, and Siyu Bao. 2024. Enhancing cross-lingual emotion detection with data augmentation and token-label mapping. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 528–533, Bangkok, Thailand. Association for Computational Linguistics.

Xiliang Zhu, Shayna Gardiner, Tere Roldán, and David Rossouw. 2024. The model arena for cross-lingual sentiment analysis: A comparative study in the era of large language models. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 141–152, Bangkok, Thailand. Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Appendix

A.1 Emotion intensity annotation agreement

Lang.	Anger	Disgust	Fear	Joy	Sadness	Surprise	Avg.
amh	0.60	0.59	0.58	0.59	0.54	0.52	**0.57**
orm	0.52	0.50	0.48	0.53	0.50	0.53	**0.51**
som	0.59	0.47	0.48	0.49	0.50	0.41	**0.49**
tir	0.55	0.54	0.52	0.51	0.53	0.53	**0.53**

Table 3: Emotion intensity Cohen's Kappa inter-annotator agreement (IAA) scores across languages

A.2 Hyperparameters

For the evaluation metrics of multi-label emotion classification, we use the Macro F1 score. For the evaluations of intensity, we used Pearson correlation scores. Fine-tuning hyperparameters of pretrained language models (PLMs) are epoch 3, lrate = 5e-5, max-token 256, and batch size 8.

LLMs prompt for the zero-shot multi-label emotion classification:

Multi-label emotion classification prompt: `"Evaluate whether the author of the following text conveys the emotion {{EMOTION}}. Think step by step before you answer. Finish your response with 'Therefore, my answer is ' ' followed by 'yes' or 'no'."`

Emotion intensity prediction prompt: `""Determine the intensity (0: none, 1: low, 2: medium, 3: high) of {{EMOTION}} in the text. "Provide reasoning and end with 'Answer:' followed by the intensity score (0..3)."`

A.3 Emotion Co-occurrence

Figure 2 shows the co-occurrence between emotion classes. Consistently in all languages, anger and disgust are the most common emotions that appear together. Anger, disgust, and joy are the top three emotions with the highest intensity level, as they also have the most statistics among other emotions, such as fear and surprise.

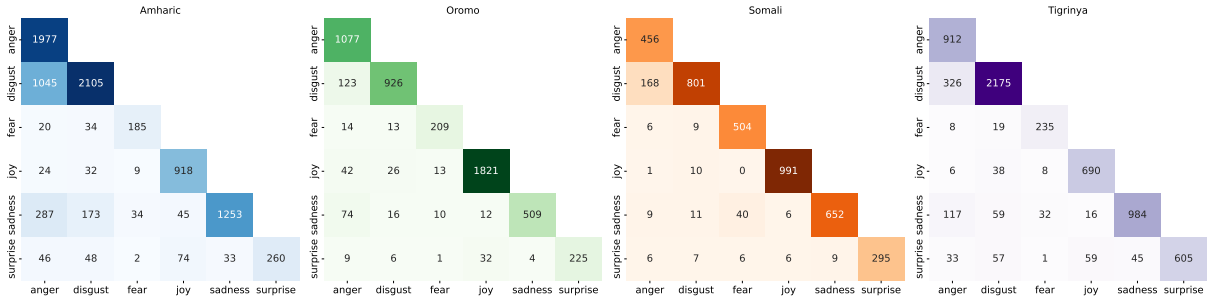

Figure 2: Emotion co-occurrence across the six basic emotions and languages

A.4 Intensity Annotation

For the intensity annotation, we employed native speakers of each language. We provided detailed annotation guidelines with text examples, emotion label(s), and each intensity level of the emotions. We compensated annotators with an hourly wage in Ethiopia. A total of 20 males and five females participated in the annotation.

A.5 Emotion label distribution

Figure 3 shows the emotion label distribution across languages. As EthioEmo is annotated in a multi-label emotion approach, a text might have no emotion, one, two, multiple, or all emotion labels; most instances across all languages have a single emotion label.

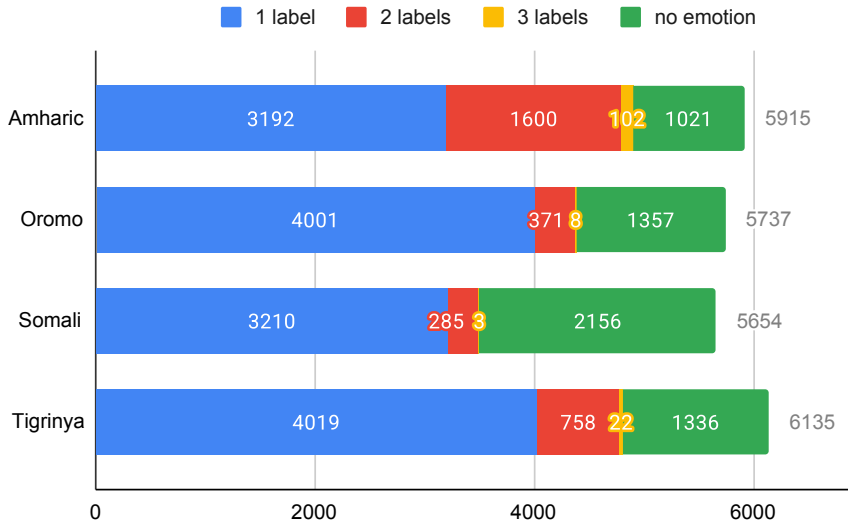

Figure 3: Emotion statistics in the number of emotion labels for each instance. Most of the instances in the dataset have a single emotion label. Amharic and Tigrinya have a comparatively high instance of two emotions for each instance.

A.6 Emotion intensity distribution

Figure 4 shows the intensity distribution.

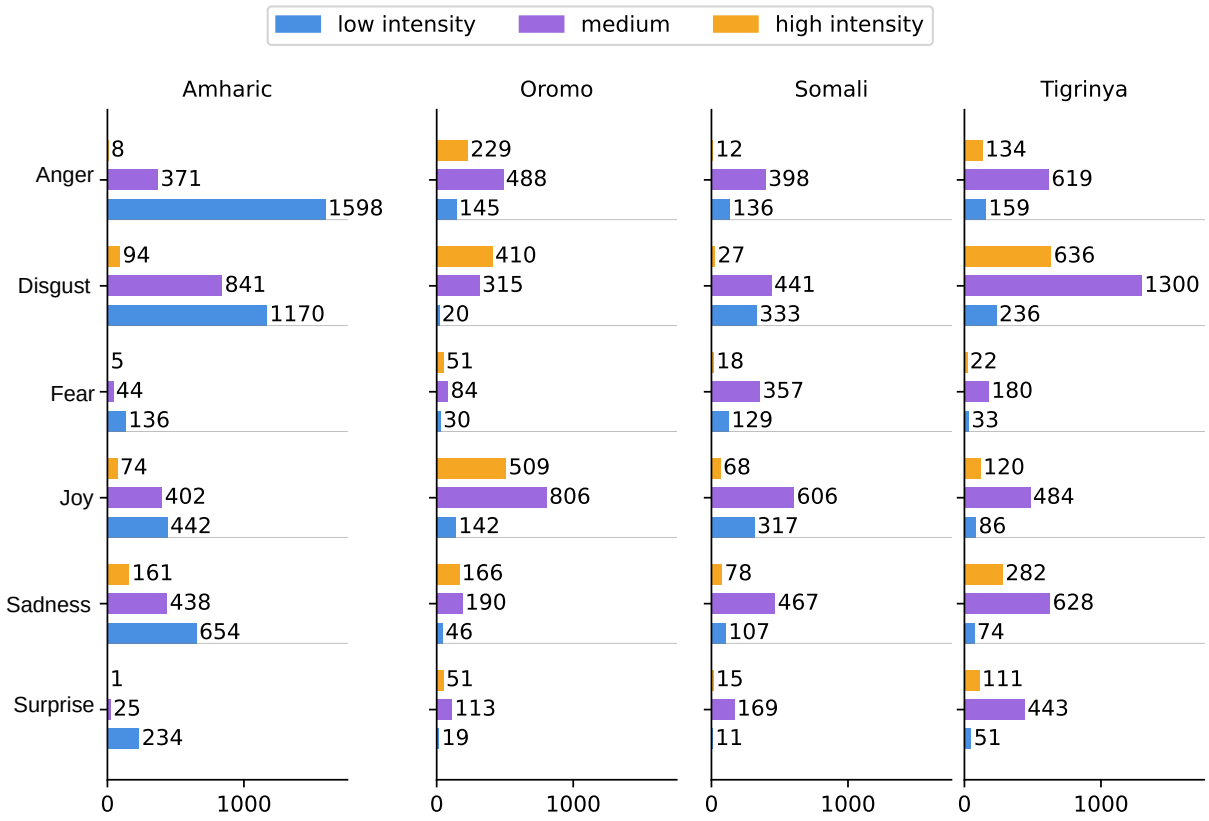

Figure 4: Emotion intensity statistics across emotion label with three intensity levels (low, medium, and high of the corresponding emotion). Instances that have not been labeled in any of the given emotions are not included in the statistics, no emotion instances are for Amharic 1021, Oromo 1357, Somali 2156, and Tigrinya 1336.

A.7 Model details and versions

- LaBSE (Feng et al., 2022) - sentence-transformers/LaBSE
- RemBERT (Chung et al., 2020) - google/rembert
- XLM-RoBERTa - FacebookAI/xlm-roberta-base (large) (Conneau et al., 2020)
- mDeBERTa (He et al., 2021) - microsoft/mdeberta-v3-base
- mBERT (Libovický et al., 2019) - google-bert_bert-base-multilingual-cased
- EthioLLM (Tonja et al., 2024) - EthioNLP/EthioLLM-l-70K : multilingual models for five Ethiopian languages (amh, gez, orm, som, and tir) and English.
- AfriBERTa (Ogueji et al., 2021) - castorini/afriberta_large : pre-trained on 11 African languages. It includes our four target Ethiopian languages.
- AfroXLM-R (Alabi et al., 2022) - Davlan/afro-xlmr-large-61L (76L) - adapted from XLM-R-large (Conneau et al., 2020) (has two versions: 61 and 76 languages) for African languages, including the four Ethiopian languages and high-resource languages such as English, French, Chinese, and Arabic.
- AfroLM (Dossou et al., 2022) - bonadossou/afrolm_active_learning - a multilingual model pre-trained on 23 African languages, including amh and orm from Ethiopian languages.
- DeepSeek-R1-70 (Guo et al., 2025) - deepseek-ai/DeepSeek-R1-Distill-Llama-70B
- Mistral-8x7B (Jiang et al., 2024) - mistralai/Mixtral-8x7B-Instruct-v0.1
- Llama-3.3-70B (Grattafiori et al., 2024) - meta-llama/Llama-3.3-70B-Instruct
- Qwen2.5-72B (Qwen et al., 2025) - Qwen/Qwen2.5-72B-Instruct
- Dolly-v2-12B (Conover et al., 2023) - databricks/dolly-v2-12b