Enhancing Multi-Label Emotion Analysis and Corresponding Intensities for Ethiopian Languages

Tadesse Destaw Belay[1], **Dawit Ketema Gete**[2], **Abinew Ali Ayele**[3], **Olga Kolesnikova**[1], **Iqra Ameer**[4], **Grigori Sidorov**[1] **and Seid Muhie Yimam**[5]
[1]Instituto Politécnico Nacional, Mexico City, Mexico, [2]Wollo University, Kombolcha, Ethiopia, [3]Bahir Dar University, Bahir Dar, Ethiopia, [4]Pennsylvania State University, PA, USA, [5]University of Hamburg, Hamburg, Germany
tadesseit@gmail.com, {tbelay23, kolesnikova, sidorov}@cic.ipn.mx, userdavek@gmail.com, abinewaliayele@gmail.com, iqa5148@psu.edu, seid.muhie.yimam@uni-hamburg.de

Abstract

Developing and integrating emotion-understanding models are essential for a wide range of human-computer interaction tasks, including customer feedback analysis, marketing research, and social media monitoring. Given that users often express multiple emotions simultaneously within a single instance, annotating emotion datasets in a multi-label format is critical for capturing this complexity. The **EthioEmo** dataset, a multilingual and multi-label emotion dataset for Ethiopian languages, lacks emotion intensity annotations, which are crucial for distinguishing varying degrees of emotion, as not all emotions are expressed with the same intensity. We extend the EthioEmo dataset to address this gap by adding emotion intensity annotations. Furthermore, we benchmark state-of-the-art encoder-only Pretrained Language Models (PLMs) and Large Language Models (LLMs) on this enriched dataset. Our results demonstrate that African-centric encoder-only models consistently outperform open-source LLMs, highlighting the importance of culturally and linguistically tailored small models in emotion understanding. Incorporating an emotion-intensity feature for multi-label emotion classification yields better performance. The data is available at https://huggingface.co/datasets/Tadesse/EthioEmo-intensities.

1. Introduction

Human emotion understanding 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 an emotion label(s) to a text that most accurately reflects the mental state of the author(writer) 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 (Machová et al., 2023; Zhang et al., 2024c; Pereira et al., 2024). How people convey their views and emotions is inherently diverse, often shaped by sociodemographic factors such as cultural background, personal experiences, communication styles, and emotional states (Hoemann et al., 2025).

It is critical to adopt systematic methodologies for organizing emotions in textual data in order to effectively analyze and interpret the complex and varied ways they are expressed. This involves employing structured annotation methods to categorize emotions and their intensity scales meaningfully (Zhang et al., 2024a). There are two approaches in annotating an emotion dataset: **single-label** and **multi-label**. In the single-label 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. Both single-label and multi-label annotations can be multiclass, i.e., datasets with more than two classes. Furthermore, emotion intensity is an extension of the emotion classification task that quantifies the strength or degree of each expressed emotion (Mashal and Asnani, 2017). In emotion annotation, especially in multi-label emotion, adding the intensity of each corresponding emotion is very crucial, as each labeled emotion might not always be equally expressed in a content (Firdaus et al., 2020; Labat et al., 2022).

Detecting the strength/intensity of emotion helps to understand its urgency and emphasis. For instance, some feelings may be subtly present, while others dominate more promi-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8056–8075
11–16 May 2026.

Figure 1: The EthioEmo dataset is a multi-label emotion dataset for Ethiopian languages, enhanced with new emotion intensity features. Each text may express one, two, several, or all emotions, and each annotated emotion is assigned an intensity level ranging from 1(low), 2(medium), or 3(high).

nently. This complexity highlights the importance of assessing intensity, as it provides a nuanced understanding of how emotions are expressed. Consider a sentence example, *'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 scale value, while others have multiple emotions, each with its own intensity scale.

The work by Belay et al. (2025b) 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 each labeled emotion. The main contributions of our works are:

- Extending the EthioEmo dataset to incorporate the multiple emotions along with their corresponding intensity to provide the complete affective information of a given instance, thereby enriching the applicability of the dataset for nuanced emotion and corresponding intensity analysis.
- Conducting a comprehensive evaluation of BERT-based encoder-only pre-trained language models (PLMs), open-source large language models (LLMs), and proprietary LLMs for their effectiveness in multi-label emotion classification and intensity prediction.
- Exploring the feasibility of cross-lingual transfer learning in a low-resource language setup and emotion transferability across the Ethiopian languages.

2. Related Work

Multi-label Emotion: Emotion is central to human nature, and as online interactions grow, users express and react to a content in various ways. A text expression is the major one and can simultaneously manifest multiple emotions to reflect the complex emotional nuances conveyed (Mashal and Asnani, 2017). To handle this complex multiple emotion expressions simultaneously, 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 Task 2 (Giorgi et al., 2024), BRIGHTER (Muhammad et al., 2025a), EthioEmo (Belay et al., 2025b), and SemEval2025 Task 11 data (Muhammad et al., 2025b) are among them. While emotion recognition is widely studied, merely identifying the type of emotion in text is often insufficient for decision-making (Al Maruf et al., 2024); its corresponding intensity is very crucial.

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

Intensity dataset	language(s)	# of isntance	Emoiton classes	Intensity lalebs
Mohammad et al. (2018)	eng,ara,spn	10,983/ 4,381/7,094	12 emotin classes	1 (low), 2 (medium), 3 (high)
Demszky et al. (2020)	eng	54.3k	27 emotion categories	No intensities
Öhman et al. (2020)	eng,fin	25k / 30k	8 emotions + neutral	No intensities
Firdaus et al. (2020)	eng	13k	6 emotions + neutral	1 (low), 2 (medium), 3 (high)
Ciobotaru et al. (2022)	eng	5,449	6 emotions + neutral	No intensities
Singh et al. (2022)	hin	1,814	15 emotion classes	1 (low), 2 (medium), 3 (high)
Rahman et al. (2024)	eng	6,037	8 depression emotions	No intensities
Muhammad et al. (2025a)	28 languages	1,645 - 9,272	6 emotion categories	1 (low), 2 (medium), 3 (high)
Plisiecki et al. (2025)	pol	10k	6 emotions	1 - 5 point scale
Belay et al. (2025b)	amh,orm,som,tir	5,915/ 5,737/ 5,654/6,135	6 emotions + neutral	1 (low), 2 (medium), 3 (high)

Table 1: Summarized multi-label emotion and intensities datasets related works.

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 scales for the corresponding labeled emotion. However, the EthioEmo dataset is annotated in a multi-label annotation setting without specifying the intensity of each corresponding emotion. Inspired by this work and the importance of incorporating intensity in multi-label emotion annotation, we extend the EthioEmo dataset by adding an intensity feature. Related work on multi-label emotion and emotion-intensity datasets is summarized in Table 1.

Cross-Lingual Experimentation: Cross-lingual transfer learning has emerged as a promising approach to address data scarcity in low-resource languages (Maladry et al., 2024). It has been used to transfer knowledge from high-resource to low-resource and among low-resource languages (Zhang et al., 2024b). 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). Cross-language experimentation could explore whether emotion classification can be improved by transferring knowledge across languages. 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 emotion analysis among Ethiopian languages, which are characterized by distinct script systems: Amharic (amh) and Tigrinya (tir) use the Ethiopic (Ge'ez) script, while Oromo (orm) and Somali (som) use the Latin script.

3. EthioEmo Dataset

The EthioEmo emotion dataset is an emotion dataset that covers four Ethiopian languages. The data was collected from social media platforms such as X(Twitter) and news portals and annotated in a multi-label setup. The targeted emotion classes are the six basic emotion classes (anger, disgust, fear, joy, sadness, and surprise). Text without an emotion label is assigned the neutral label "0". Each instance of the dataset is annotated by a minimum of three and a maximum of five annotators. The final emotion labels are determined through a majority vote (two or more votes for the three annotators and three or more votes for the five annotators).

Emotion Intensity Annotation The degree of feeling in an emotion dataset is crucial for accurately understanding complex emotions (Firdaus et al., 2020). We enhanced the EthioEmo dataset by including annotations for the intensity of each identified emotion. Annotators were trained to assign an intensity label to each emotion category. We utilized the customized version of the POTATO annotation tool (Pei et al., 2022) along with in-house annotation practices. Annotators are trained to assign an intensity label to each identified emotion category. 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 com-

prises four levels: 0 (no intensity for any emotion class — neutral), 1 (slight emotion, e.g., slight anger), 2 (moderate emotion, e.g., moderate anger), and 3 (high emotion, e.g., very anger). Following the original dataset's annotation setup and based on annotator availability, each instance in `orm`, `som`, and `tir` was annotated by a minimum of three annotators, while each instance in `amh` is annotated by five annotators, and the final label was determined by majority vote. The rationale for assigning a minimum of three annotators is based on previous annotations of related datasets (Belay et al., 2025b; Mohammad et al., 2018; Muhammad et al., 2025b). It also represents the minimum cost-effective number required to obtain a majority vote, particularly in settings with limited annotator availability. Annotation guidelines and details of annotators are found in Appendix G.

The final intensity score for each emotion is obtained using the following aggregation rule. For three annotators per instance, at least two annotators must assign a non-zero intensity value (1, 2, or 3); instances with fewer than two non-zero annotations are discarded, see rule 1. In the five-annotator setting (rule 2), the rule is applied whenever at least two annotators assign a non-zero intensity. The intensity of each emotion for the three- and five-annotator settings is then computed using the following formulas.

For three annotators per instance (1):

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

For five annotators per instance (2):

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

where Avg is the average intensity score among the annotators of an instance.

Annotators Agreement We obtained moderate results of inter-annotator agreement (IAA) based on Cohen's Kappa (see Table 2), where scores < 0.0 indicate poor agreement, 0.00–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect (Sánchez-Velázquez and Sierra, 2016). The moderate intensity IAA agreement score shows the difficulty of the emotion's intensity annotation task.

Lang.	Ang.	Disg.	Fear	Joy	Sadn.	Surp.	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 2: Emotion intensity Cohen's Kappa (Cohen, 1960) inter-annotator agreement (IAA) scores across languages and emotions.

Emotion Intensity Distribution The distribution of the emotions and intensities in the dataset is presented in Figure 2. Most instances have low or medium emotion intensity, while `Joy` in `orm`, `Disgust` in `tir`, and `orm` languages have many instances with high emotion intensity compared to other emotion classes.

4. Experiment Setup

4.1. Model Selection

We select language models for evaluation from different perspectives, such as general multilingual PLMs, Africa-centric PLMs, open-source LLMs, and proprietary LLMs. The rationale behind choosing LLMs is that small and large variants (such as Llama-3.1-8B and Llama-3.3-70B) focus on multilingual support and popularity.

General Multilingual PLMs We evaluate the most common multilingual BERT-like PLMs such as LaBSE (Feng et al., 2022), RemBERT (Chung et al., 2021), 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),

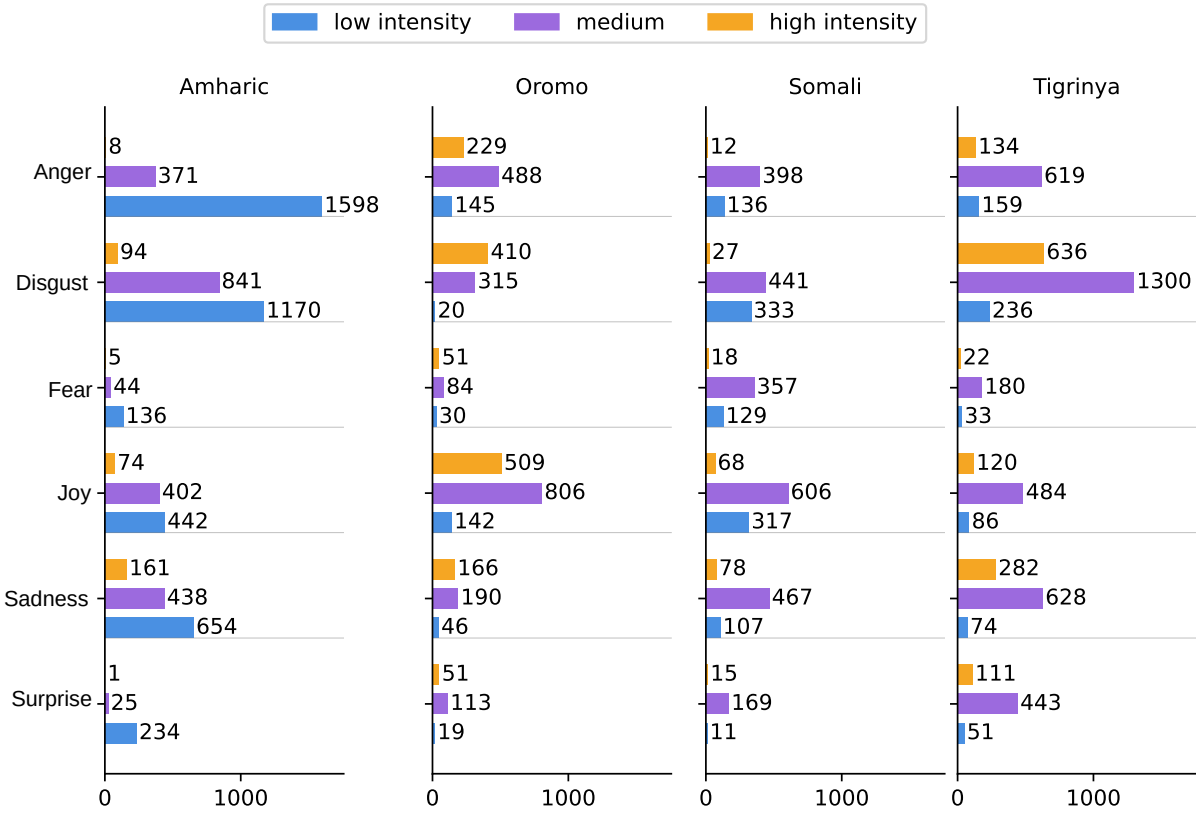

Figure 2: Emotion intensity distribution across emotion labels with three intensity levels (low, medium, and high of the corresponding emotion). Instances that have not been labeled in any of the given emotions are not included in the statistics, instances that have none of the targeted emotions are for Amharic (amh) is 1021, Oromo (orm) 1357, Somali (som) 2156, and Tigrinya (tir) 1336. The statistics of each emotion across all languages can be found in Appendix D.

AfroXLMR (61 and 76 languages) (Alabi et al., 2022), EthioLLM (Tonja et al., 2024), and AfroXLMR-Social (Belay et al., 2025a).

Open source LLMs Based on their popularity in the open-source community and better multilingual support, we evaluate the following open-source LLMs: Gemma-3-12B (Kamath et al., 2025), Llama-3.1-8B (Dubey et al., 2024), Llama-3.3-70B (Dubey et al., 2024), and DeepSeek-R1-70B (Guo et al., 2025).

Proprietary LLMs We examine the lightweight and latest versions of the proprietary models, GPT-4.1-mini (OpenAI et al., 2025) and Gemini-2.5-pro-flash (DeepMind, 2025) for reproducibility and cost-effectiveness.

4.2. Evaluation Setup

For encoder-only models, we fine-tune them using the train–test split of the dataset for emotion classification, emotion intensity prediction, and cross-lingual transfer experiments. The training experiment settings are presented in Appendix A for reproducibility. For LLMs, Chain-of-Thought (CoT) prompting is a widely used strategy, particularly effective across various NLP tasks, including mathematical reasoning. However, prompting models to select from a predefined list of emotions, such as prompting LLMs to choose the emotion of a given text from [anger, disgust, fear, joy, sadness, surprise], often leads to over-prediction, where models assign multiple emotion classes even when most instances contain only a single emotion. To mitigate this, we employed CoT prompting (Wei et al., 2022) with binary emotion prediction, evaluating one emotion at

a time with a yes/no response.

4.3. Formulating Evaluation Tasks

Using the aforementioned language models and the multi-label emotion dataset with corresponding intensity scales, we conduct experiments with a 60:10:30 train–dev–test split from a total datsets amh - 5,915, orm – 5,737, som – 5,654, and *tir* – 6,135. Using these datasets, we evaluate the following emotion analysis tasks:

- **Multi-label emotion classification**: The emotion classification task is formulated as a binary decision for each emotion (one emotion at a time), where the model provides a yes/no (1/0) response to questions such as "*Does the text convey anger or not?*" The same setup is applied for all other emotions.

- **Emotion intensity prediction**: For intensity prediction, we provide the text along with list of emotions and ask the model to predict the intensity level of each corresponding emotion from 0 (no), 1(low), 2 (medium), or 3 (high).

- **Cross-lingual multi-label emotion evaluation**: Our cross-lingual experiment setup involves two types: 1) fine-tuning BERT-like encoder-only models on datasets from all available languages except the target language, which is held out for evaluation, and 2) fine-tuning a single multilingual model that includes all the languages.

4.4. Evaluation Metrics

Based on the multi-label task evaluations from previous similar works (Mohammad et al., 2018; Muhammad et al., 2025b) and for a highly imbalanced data, we used the Macro-F1 score that takes the average of individual emotion F1 scores. For intensity prediction, we used the Pearson correlation coefficient (*r*) (Schober et al., 2018) between predicted and true intensity values.

5. Experiment Results

5.1. Multi-Label Emotion Classification

The results of the multi-label emotion classification are presented in Table 3. As shown, BERT-like encoder-only models achieve better performance than zero-shot LLMs. AfroXLMR-Social achieves stronger results, possibly due to three main reasons: (1) it is a continual pretarined model from African-centric AfroXLMR model using domain specific social media corpus - a corpus from X (Twitter) and news, (2) it is based on multilingual XLM-RoBERTa that covers 100 languages (Conneau et al., 2020) and further fine-tuned on 76 African languages, and (3) it benefits from a larger parameter size (comparatively across encoder-only models) and more diverse training data than EthioLLM. AfroXLMR, the base of AfroXLMR-Social, was trained on approximately ≈ 12 GB of multilingual African text, which enables more effective cross-lingual transfer than EthioLLM (trained on ≈ 3 GB corpus).

Large language models (LLMs) perform less effectively for the evaluated low-resource languages, and their performance is highly dependent on parameter size. For example, Llama-3.1-8B performs the worst among evaluated LLMs, while Llama-3.3-70B performs better. Comparably, Amharic is better represented than other Ethiopian languages. Overall, encoder-only models continue to outperform both open-source and proprietary LLMs in the multi-label emotion analysis for low-resource languages. African-centric PLMs are better at classifying multi-label emotions than LLMs.

How does adding corresponding intensity features enhance multi-label emotion performance? We evaluate our best encoder-only model, AfroXLMR-Social, to assess the impact of incorporating an emotion intensity feature for multi-label emotion classification. The model achieves macro F1 scores of 82.13 for amh, 62.36 for orm, 57.53 for som, and 65.19 for tir language. These results represent improvements ranging from 1.62 to 11.47 points compared to the baseline multi-label emotion-only results that are shown in Table 3. Training with emotion intensity annotations provides a

Models	Multi-Label Emotion Classification (F1)					Intensity Prediction (Pearson *r*)				
	amh	orm	som	tir	Avg.	amh	orm	som	tir	Avg.
General multilingual PLMs										
LaBSE	66.51	41.49	43.99	48.88	50.22	47.79	16.53	25.70	32.10	30.53
RemBERT	60.15	47.54	48.31	50.37	51.59	52.73	24.15	24.85	37.63	34.84
mBERT	26.51	40.32	27.01	25.72	29.89	00.00	17.88	5.51	3.13	6.63
mDeBERTa	53.43	32.84	36.86	41.73	41.22	33.07	7.27	7.02	19.24	16.15
XLM-RoBERTa	63.73	37.42	33.51	13.32	37.00	53.63	17.34	18.39	15.95	26.33
African-centric PLMs										
EthioLLM	58.68	47.95	33.84	44.78	46.31	41.90	21.58	9.96	22.77	24.05
AfriBERTa	60.64	54.10	44.66	47.97	53.34	39.38	25.24	20.63	27.56	28.20
AfroLM	54.76	42.21	32.77	38.60	42.09	37.75	15.90	5.08	18.42	19.25
AfroXLMR-61L	67.93	51.73	49.31	54.96	55.98	55.19	26.75	37.81	41.96	40.43
AfroXLMR-76L	68.46	49.68	49.25	53.08	55.12	**55.57**	29.15	41.36	40.32	41.60
AfroXLMR-Social	**70.66**	**60.74**	**54.75**	**60.24**	**61.60**	53.82	**32.26**	**38.44**	**42.18**	**41.68**
Zero-shot - open-source LLMs										
Llama-3.1-8B	32.06	07.77	07.13	11.84	14.46	14.58	07.13	09.36	07.30	09.59
Gemma-3-12B	42.19	23.28	32.05	32.57	32.62	30.45	16.60	26.08	**22.17**	23.83
DeepSeek-R1-70B	36.89	28.15	26.56	26.49	29.52	31.05	25.17	26.26	21.78	26.07
Llama-3.3-70B	**42.84**	**29.84**	**32.49**	**32.93**	**34.53**	**39.52**	**27.31**	**30.08**	21.12	**29.51**
Zero-shot - commercial LLMs										
Gemini-2.5-flash	24.56	24.11	16.38	12.31	19.34	24.51	14.08	11.16	9.46	14.80
GPT-4.1-mini(0-shot)	**46.73**	44.06	45.07	**34.77**	42.66	40.01	37.20	41.49	**29.71**	37.35
GPT-4.1-mini(5-shot)	45.68	**46.60**	**48.10**	32.73	**43.28**	**41.21**	**37.59**	**45.94**	28.85	**38.40**

Table 3: Multi-label emotion prediction macro F1 results (left) and emotion intensity prediction (right) results using Pearson correlation. The best performance scores are highlighted in **bold**. All evaluated open-source LLMs are instructed versions.

denser supervision signal than binary emotion-only labels. In particular, the model learns to emphasize high-intensity emotional signals while naturally reducing the ambiguity associated with low-intensity or borderline cases.

5.2. Emotion Intensity Prediction

Table 4 presents the results of the intensity prediction. As all Ethiopian languages are not included during pretraining, mBERT performs worse; the slightly better performance on `orm` and `som` might be because these languages use the Latin script and share some vocabulary. Similarly, in the emotion classification task, AfroXLMR-Social achieves the highest performance in intensity prediction. LLMs at intensity prediction are worse than the emotion classification task, this might be due to the subjectivity and complexity of emotion intensity prediction pose a greater challenge even for high-resource languages, such as English (Muhammad et al., 2025a). The overall results show that understanding emotions and predicting intensities from text is challenging.

How do LLMs help for emotion and its intensity annotation for low-resource languages? State-of-the-art LLMs, such as GPT, demonstrate close to human-level performance in generating high-quality emotion and intensity annotations for English (Pavlovic and Poesio, 2024; Bagdon et al., 2024). However, for low-resource languages, their performance drops significantly - below 50% for all Ethiopian languages . A closer examination of the GPT prediction outputs reveals that the model often attempts to translate the input text into English before predicting emotions and their corresponding intensities (particularly for Ethiopic script languages such as `amh` and `tir`), despite explicit prompt instructions to avoid explanations or translations. Consequently, the model frequently outputs “no emotion” and “no intensity” for these languages.

Models	Cross-lingual results (F1)					Cross-lingual with same script (F1)				
	amh	orm	som	tir	Avg.	amh	orm	som	tir	Avg.
LaBSE	44.11	20.77	35.18	40.13	35.55	43.83	21.60	21.09	36.99	30.88
RemBERT	42.65	20.87	31.32	33.39	31.81	37.81	33.90	23.11	12.31	26.78
mBERT	25.10	10.79	14.13	18.27	17.07	28.15	23.51	18.10	19.95	22.43
mDeBERTa	36.40	26.63	18.83	38.03	29.97	38.23	24.53	13.92	33.05	27.43
XLM-RoBERTa	23.52	23.69	26.98	38.63	28.21	31.65	21.63	09.22	21.56	21.02
EthioLLM	38.37	22.46	22.76	33.08	30.42	31.31	20.90	10.08	27.40	22.42
AfriBERTa	46.28	35.86	30.81	38.05	37.75	38.02	27.09	26.75	31.37	30.81
AfroLM	32.12	10.38	9.00	25.48	19.25	28.67	23.05	11.53	21.75	21.25
AfroXLMR-61L	56.41	43.24	42.21	52.70	48.64	53.13	27.84	12.68	44.57	34.56
AfroXLMR-76L	56.65	45.01	41.24	53.39	49.07	45.95	28.47	14.06	49.65	34.53
AfroXLML-Social	**60.22**	**52.30**	**45.72**	**56.00**	**53.56**	**59.28**	**41.44**	**37.00**	**53.65**	**47.84**

Table 4: Cross-lingual emotion prediction results. Train with all languages except the held-out language results (left) and train with only similar scrips (right) column. The best performance are highlighted in **bold**.

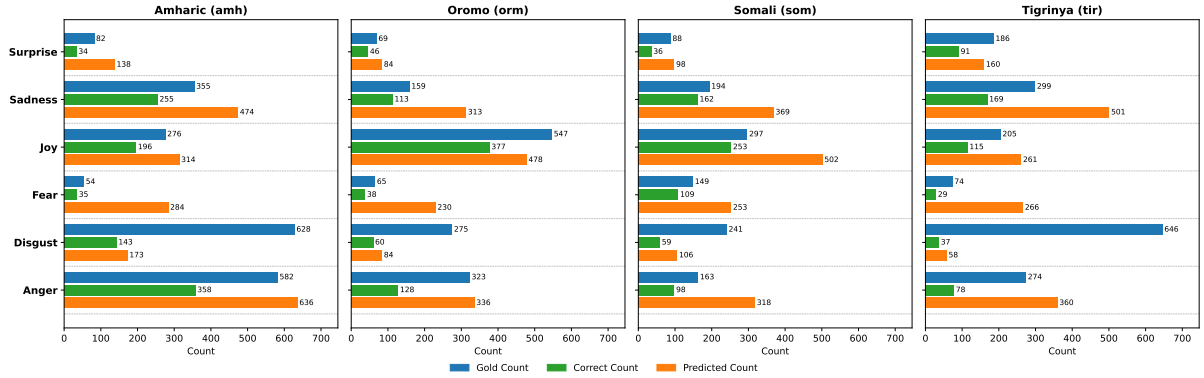

Figure 3: Emotion error analysis across languages and Emotion labels from the better LLMs (GPT-4.1-mini 5-shot). **Gold Count** is the number of human annotated labels, **Correct Count** is the labels the model predicted correctly from the total given gold count, and **Predicted Count** is the total predictions of the model for that specific emotion.

5.3. Cross-lingual Emotion Classification

How does cross-lingual emotion transferability work among Ethiopian languages? Table 4 reports the results of the cross-lingual transfer learning experiments. AfroXLMR-Social again achieves the highest performance for cross-lingual evaluations because it includes all the targeted Ethiopian languages during pretraining. Overall, when comparing cross-lingual results across **all multilingual** and **only similar-script** settings, all languages benefited from multilingual training. Specifically, Amharic (amh) and Tigrinya (tir) achieved higher transfer performance, likely because both use the Ethiopic script (Ge'ez). Among the BERT-like models, mBERT performs the worst, as none of the target languages was included during its pretraining. The AfroLM model performs the second worst, as it includes only Amharic (amh) in its pretraining.

Error Analysis We conducted a detailed prediction analysis for both emotion and intensity prediction to understand why LLMs (specifically GPT-4.1-mini 5-shot in this analysis) perform worse than BERT-like models. Figure 3 presents detailed statistics, including the number of human-annotated labels (Gold Count), the number of correctly predicted labels (Cor-

rect Count) out of the total gold labels and the total number of predictions made by the model for each emotion (Predicted Count). Based on the statistics: 1) The number of correctly predicted labels is consistently lower than the number of gold labels, and 2) Mostly the model tends to over-predict, except for the `Disgust` emotion class across language. Although the disgust emotion has one of the highest distributions across languages in the EthioEmo dataset, it is predicted less frequently than other emotion classes. A similar trend is observed in the intensity prediction results, where most errors are over-predictions, for example, assigning emotion intensities to instances that are not labeled with any emotion, and predicting 2 (medium) and 3 (high) intensities for instances that have low intensity annotations.

6. Conclusion

In this work, we extended 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 such as AfroXLMR-Social that includes the evolution languages during pretraining performs is best for emotion, intensity, and cross-lingual emotion transferability between Ethiopian languages. This dataset and benchmark will contribute to the development of a more robust emotion evaluation task for low-resource languages. In future work, we plan to release the annotator level data and 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 for subjective NLP tasks such emotion analysis and emotion intensity prediction.

Limitations

Our work is not without limitation and we identified the following limitations with the future directions.

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, and `amh` has five annotators per instance. Future work can add more annotations on top of our three annotators for the more 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 the annotator-level data publicly available and open it for further exploration to determine the final emotion intensity. Alternatively, it can be used to model annotator perspectives without applying a majority vote.

Limited models 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, closed-source LLMs, and various few-shot evaluation setup. Additionally, the data can be explored in many ways such as if certain emotions or languages are more subjective than others.

Ethical Considerations

As started from a previously published dataset (Belay et al., 2025b), 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.

7. Bibliographical References

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

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.

Christopher Bagdon, Prathamesh Karmalkar, Harsha Gurulingappa, and Roman Klinger. 2024. "You are an expert annotator": Automatic Best–Worst-Scaling Annotations for Emotion Intensity Modeling". In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 7924–7936, Mexico City, Mexico. Association for Computational Linguistics.

Tadesse Destaw Belay, Israel Abebe Azime, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Idris Abdulmumin, Abinew Ali Ayele, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025a. AfroXLMR-Social: Adapting Pre-trained Language Modelsfor African Languages Social Media Text. *arXiv preprint arXiv:2503.18247*.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025b. 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 Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking Embedding Coupling in Pre-trained Language Models. In *International Conference on Learning Representations*.

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.

Jacob Cohen. 1960. A coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1):37–46.

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.

DeepMind. 2025. Gemini 2.5: Our most intelligent AI model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. [Accessed 15-10-2025].

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, pages arXiv–2407.

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*, 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 Pre-Training with Gradient-Disentangled Embedding Sharing.

Katie Hoemann, Yeasle Lee, Èvelyne Dussault, Simon Devylder, Lyle H Ungar, Dirk Geeraerts, and Batja Mesquita. 2025. The construction of emotional meaning in language. *Communications Psychology*, 3(1):99.

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.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

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?

Kristína Machová, Martina Szabóova, Ján Paralič, and Ján Mičko. 2023. Detection of emotion by text analysis using machine learning. *Frontiers in Psychology*, Volume 14 - 2023.

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, LA, 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 Tufiño, 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, Alexander Panchenko, Andrew Piper, 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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 8895–8916, Vienna, Austria. Association for Computational Linguistics.

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.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552, Barcelona, Spain (Online). International Committee on Computational Linguistics.

OpenAI, Ananya Kumar, Juahui Yu, John Hallman, and Michelle Pokrass et al. 2025. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. Accessed: 2025-10-22.

Maja Pavlovic and Massimo Poesio. 2024. The Effectiveness of LLMs as Annotators: A Comparative Overview and Empirical Analysis of Direct Representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The Portable Text Annotation Tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Rafael Pereira, Carla Mendes, Jose Ribeiro, Roberto Ribeiro, Rolando Miragaia, Nuno Rodrigues, Nuno Costa, and António Pereira. 2024. Systematic review of emotion detection with computer vision and deep learning. *Sensors*, 24(11):3484.

Hubert Plisiecki, Piotr Koc, Maria Flakus, and Artur Pokropek. 2025. Predicting emotion intensity in polish political texts: comparing supervised models and large language models in a low-resource language. *Quality & Quantity*, 59:3405—3427.

Abu Bakar Siddiqur Rahman, Hoang-Thang Ta, Lotfollah Najjar, Azad Azadmanesh, and Ali Saffet Gönul. 2024. DepressionEmo: A novel dataset for multilabel classification of depression emotions. *Journal of Affective Disorders*, 366:445–458.

Octavio Sánchez-Velázquez and Gerardo E Sierra. 2016. Let's Agree to Disagree: Measuring Agreement between Annotators for Opinion Mining Task. *Research on computing science*, 110:9–19.

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation Coefficients: Appropriate use and Interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Fa Zhang, Jian Chen, Qian Tang, and Yan Tian. 2024a. Evaluation of emotion classification schemes in social media text: an annotation-based approach. *BMC psychology*, 12(1):503.

Jinghui Zhang, Yuan Zhao, Siqin Zhang, Ruijing Zhao, and Siyu Bao. 2024b. 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.

Junbo Zhang, Qi Chen, Jiandong Lu, Xiaolei Wang, Luning Liu, and Yuqiang Feng. 2024c. Emotional expression by artificial intelligence chatbots to improve customer satisfaction: Underlying mechanism and boundary conditions. *Tourism Management*, 100:104835.

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.

Appendix

A. Model Training/Testing Parameters

Fine-tuning hyperparameters of encoder-only PLMs are epoch 3, lrate = 5e-5, max-token 256, and batch size 8.

Multi-label emotion classification prompt: `"Evaluate whether the author of the following text conveys the emotion {{EMOTION}}. Think step by step before you answer. Answer with ONLY 'yes' or 'no'. Do not provide any explanation.`
`Text: {text} "`

Emotion intensity prediction prompt: `Determine the intensity of the emotions {{EMOTION}} in the text. The intensity score ranges from 0 to 3:`
`- 0 = No intensity (emotion not present)`
`- 1 = Low intensity`
`- 2 = Medium intensity`
`- 3 = High intensity`
`Example output:`
`{{"anger": 0, "disgust": 2, "fear": 0, "joy": 1, "sadness": 0, "surprise": 0}}`
`Text: {text}`

B. Additional Results

Table 5 presents detail results at each emotion class level from the best AfroXLMR-Social (Belay et al., 2025a) model.

	Emotion results (F1)				Intensity results (F1)			
Models	amh	orm	som	tir	amh	orm	som	tir
Anger	70.29	56.97	48.29	41.32	47.35	27.98	13.81	00.00
Disgust	79.37	64.21	55.31	77.46	69.72	48.60	38.50	67.18
Fear	57.78	38.25	56.30	43.52	00.00	00.00	53.60	00.00
Joy	75.91	82.64	66.85	64.66	78.58	67.71	57.81	59.32
Sadness	73.30	56.12	68.43	65.79	80.16	49.30	66.91	60.23
Surprise	67.34	66.22	33.33	68.67	47.63	00.00	00.00	66.32
Average	70.66	60.74	54.75	60.24	53.82	32.26	38.44	42.18

Table 5: Emotion class-level results from the best model (AfroXLML-Social).

C. Emotion Co-occurrence

Figure 4 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.

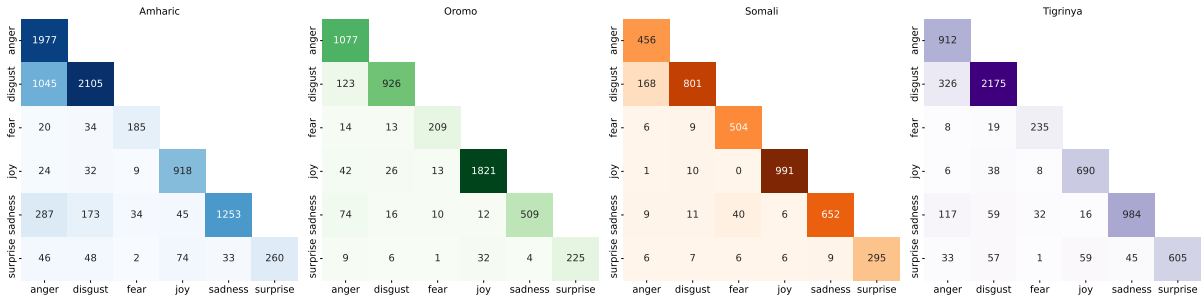

Figure 4: Emotion co-occurrence across the six basic emotions and languages

D. Emotion label distribution

Figure 5 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.

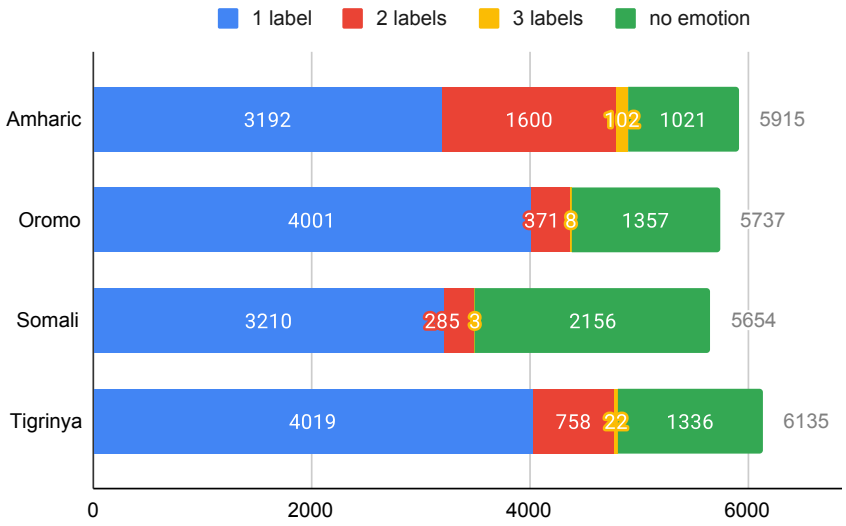

Figure 5: Emotion statistics in the number of emotion labels for each instance. Most of the instances in the dataset have a single emotion label. Amharic (1600) and Tigrinya (758) have a comparatively higher instance of double emotions than Oromo (371) and Somali (285).

E. Emotion intensities prediction analysis

Table 3 shows the prediction distributions from GPT (GPT-4.1-mini 5-shot) intensity scales from 0 (none) to 3 (high) for the corresponding emotion and across languages.

Emotion	Intensity	Amharic			Oromo			Somali			Tigrinya		
		Gold	Correct	Pred	Gold	Correct	Pred	Gold	Correct	Pred	Gold	Correct	Pred
Anger	0	1674	819	982	1436	788	853	1364	703	774	1566	493	535
	1	54	23	723	41	18	447	41	16	456	49	18	651
	2	166	47	326	98	30	322	147	37	355	179	39	332
	3	41	12	307	31	13	283	25	7	264	46	15	322
Disgust	0	1288	1043	1348	1139	1004	1308	1212	981	1226	1194	946	1304
	1	69	22	410	72	23	414	65	24	394	68	21	409
	2	391	45	120	387	56	129	395	59	112	398	47	118
	3	171	6	9	162	9	8	176	8	8	180	7	9
Fear	0	1690	1201	1192	1750	1153	1225	1774	1174	1186	1766	1162	1195
	1	12	3	366	14	3	347	12	3	355	10	2	366
	2	67	15	229	56	13	237	61	17	241	57	14	231
	3	8	3	49	6	2	44	8	3	45	7	4	48
Joy	0	1588	1334	1409	1689	1351	1447	1627	1367	1389	1635	1345	1413
	1	25	6	179	27	7	183	24	6	172	26	5	173
	2	136	30	167	129	32	175	138	33	168	134	28	169
	3	41	16	86	43	15	82	44	17	88	45	15	85
Sadness	0	1495	704	724	1534	694	722	1527	671	732	1541	655	721
	1	16	5	530	17	4	527	20	6	544	18	5	540
	2	185	49	389	191	50	394	189	48	388	192	50	396
	3	93	32	184	84	29	183	87	27	186	89	31	183
Surprise	0	1662	1437	1503	1655	1433	1497	1648	1441	1516	1654	1429	1502
	1	12	4	284	13	5	281	10	4	277	11	4	283
	2	135	32	57	138	30	58	136	33	55	137	31	55
	3	39	0	0	36	0	0	37	0	0	38	0	0

Table 6: Emotion intensity distribution across languages. The intensity levels are defined as 0 = no emotion, 1 = low, 2 = medium, and 3 = high for each corresponding emotion category. **Gold** denotes the number of human-annotated instances, **Correct** indicates the number of instances the model correctly predicted among the Gold labels, and **Pred.** (Prediction) represents the total number of instances the model predicted for that specific intensity level. Results are obtained from the best-performing LLM, GPT-4.1-mini, evaluated in a 5-shot setting.

F. Model details, papers, and its Hugging-face name

- LaBSE (Feng et al., 2022) - sentence-transformers/LaBSE
- RemBERT (Chung et al., 2021) - 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.
- AfroXLMR (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.
- AfroXLMR-Social (Belay et al., 2025a) - Tadesse/AfroXLMR-Social - 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
- Gemma-3-12B (Kamath et al., 2025) - google/gemma-3-12b-it
- Llama-3.1-8B (Dubey et al., 2024) - meta-llama/Llama-3.1-8B-Instruct
- Llama-3.3-70B (Grattafiori et al., 2024) - meta-llama/Llama-3.3-70B-Instruct

G. Annotation Guideline

Based on previous emotion and its corresponding intensity annotation guidelines for other languages, we prepared intensity annotation guidelines. We asked annotators as which of the options below best describes the feelings of the narrator (select one option for each row): 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. Their academic status is a bachelor's degree or above.

0: No anger	1: slight anger	2: moderate anger	3: high anger
0: No disgust	1: slight disgust	2: moderate disgust	3: high disgust
0: No sadness	1: slight sadness	2: moderate sadness	3: high sadness
0: No fear	1: slight fear	2: moderate fear	3: high fear
0: No joy	1: slight joy	2: moderate joy	3: high joy
0: No surprise	1: slight surprise	2: moderate surprise	3: high surprise

G.1. Emotion Categories and Definitions

Joy: Expressions of happiness, pleasure, or contentment. Consider happiness to be a broad category that includes: joyful, elated, content, cheerful, blissful, delighted, gleeful, satisfied, ecstatic, upbeat, pleased, etc. "I just passed my exams! "

Sadness: Expressions of unhappiness, sorrow, or disappointment. Consider sadness to be a broad category that includes: melancholic, despondent, gloomy, heartbroken, longing, mourning, dejected, downcast, disheartened, dismayed, etc. "I miss my family so much. It's been a tough year."

Anger: Expressions of frustration, irritation, or rage. Consider anger to be a broad category that includes: irritated, annoyed, aggravated, indignant, resentful, offended, exasperated, livid, irate. etc. "Why is the internet so slow today?!"

Fear: Expressions of anxiety, apprehension, or dread. Consider fear to be a broad category that includes: frightened, alarmed, apprehensive, intimidated, panicky, wary, dreadful, shaken, etc. "There's a huge storm coming our way. I hope everyone stays safe."

Surprise: Expressions of astonishment or unexpected events. Consider surprise to be a broad category that includes: taken aback, bewildered, astonished, amazed, startled, stunned, taken aback, shocked, dumbstruck, confounded, stupefied, etc. "I can't believe he just proposed to me!"

Disgust: A reaction to something offensive or unpleasant. Consider disgust to include distinguishing an individual/organization based solely on their identity/humanity, i.e., religion, ethnicity, language, and insulting, belittling, or obscene words. It means hating a person for his humanity. "I hate black people"

Summary Instructions Carefully read the detailed instructions before proceeding with the task. You will be given an language sentence taken randomly from a social media (X, video comments, and news headlines) and multiple choice options for the emotions the narrator is feeling. Mark all options that apply. The options correspond to seven emotions (anger, sadness, fear, disgust, happiness, and surprise) and select the intensity of the chosen emotion on a scale from no emotion, slight emotion, moderate emotion, to high emotion.

Quality Control Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will give you immediate feedback in a pop-up box. An occasional misanswer is okay. However, if the rate of misanswering is high (e.g., >20%), then all of one's HITs may be rejected.

G.2. Additional Results from LLMs

Table 7 shows additional LLMs results from 0-shot and 5-shot. Across all tested languages, the Gemma-3 models consistently outperform the competition, with the 12b variant achieving the highest scores in both 0-shot and 5-shot settings. While most models show performance gains when moving from 0-shot to 5-shot prompts, others, such as LLaMa-3.1 and Ministral-8b, occasionally exhibit performance degradation or stagnation, suggesting a struggle with in-context learning in these low-resource languages.

	amh		orm		som		tir	
Model	0 shot	5 shot	0 shot	5 shot	0 shot	5 shot	0 shot	5 shot
Gemma-3-12b-it	52.94	53.08	32.48	36.77	41.00	46.60	41.78	39.39
Gemma-3-4b-it	45.31	47.88	19.78	22.35	33.60	38.35	30.81	36.67
LLaMa-3.1-8b-instruct	26.00	20.41	19.55	20.86	20.28	21.56	14.52	14.12
LLaMa-3.2-3b-instruct	6.61	15.56	5.85	12.19	9.94	13.59	4.71	12.27
Ministral-3b-instruct	0.00	0.37	0.00	6.19	0.00	6.08	0.00	0.00
Ministral-8b-instruct-2410	16.90	10.81	16.16	14.96	14.26	14.84	12.40	9.85
Qwen2.5-3B-Instruct	10.93	11.00	10.21	10.32	11.51	11.22	10.17	12.73
Qwen2.5-7B-Instruct	16.32	21.24	1.68	15.03	2.45	20.59	14.40	16.80

Table 7: Additional results from the LLMs' 0-shot and 5-shot.