Evaluating the Capabilities of Large Language Models for Multi-label Emotion Understanding

Tadesse Destaw Belay[1,2,*], **Israel Abebe Azime**[3,*], **Abinew Ali Ayele**[4,5], **Grigori Sidorov**[1], **Dietrich Klakow**[3], **Philipp Slusallek**[3], **Olga Kolesnikova**[1], **Seid Muhie Yimam**[5]

[1]Instituto Politécnico Nacional (IPN), CIC, [2]Wollo University, [3]Saarland University, [4]Bahir Dar University, [5]University of Hamburg

Abstract

Large Language Models (LLMs) show promising learning and reasoning abilities. Compared to other NLP tasks, multilingual and multi-label emotion evaluation tasks are under-explored in LLMs. In this paper, we present **EthioEmo**, a multi-label emotion classification dataset for four Ethiopian languages, namely, Amharic (amh), Afan Oromo (orm), Somali (som), and Tigrinya (tir). We perform extensive experiments with an additional English multi-label emotion dataset from SemEval 2018 Task 1. Our evaluation includes encoder-only, encoder-decoder, and decoder-only language models. We compare zero and few-shot approaches of LLMs to fine-tuning smaller language models. The results show that accurate multi-label emotion classification is still insufficient even for high-resource languages such as English, and there is a large gap between the performance of high-resource and low-resource languages. The results also show varying performance levels depending on the language and model type. **EthioEmo** is available publicly[1] to further improve the understanding of emotions in language models and how people convey emotions through various languages.

1 Introduction

In today's digital age, individuals freely express their feelings, arguments, opinions, and attitudes on websites, micro-blogs, and social media platforms. This situation has increased interest in extracting user sentiments and emotions towards events for various purposes such as decision-making, product analysis, customer feedback analysis, political promotions, marketing research, and social media monitoring (Kusal et al., 2022).

Emotion classification is one of the most challenging NLP tasks, where a given text is assigned to the most appropriate emotion(s) that best reflect(s) the author's mental state (Tao and Fang, 2020). It poses more challenges than similar NLP tasks, such as sentiment analysis. The challenges of emotion classification as a task worth exploring include many classes, the possibility of a single text expressing multiple emotions, and the cultural and language differences inherent in interpreting or transferring emotions (Kusal et al., 2023; Wang et al., 2024b).

Multi-label Emotion Classification (MLEC) considers all emotions expressed in a text, which is a more challenging but essential NLP task, as a text can express multiple emotions simultaneously (Ameer et al., 2020; Deng and Ren, 2020). Multi-label classification enables an instance to have any combination (none, one, some, or all) of labels from a given set of emotions.

This work intends to create and evaluate a multi-label text emotion dataset for the following Ethiopian languages: Amharic (amh), Afan Oromo (orm), Somali (som), and Tigrinya (tir), with an available English dataset for evaluation. As it makes the task more intricate and reflects the complexity often found in real-world data (Liu et al., 2023), we follow the multi-label emotion classification approach.

The main contributions are summarized as:

1. We introduce EthioEmo, a new multi-label emotion benchmark dataset for four Ethiopian languages.
2. We explore popular Afri-centric encoder-only models that include most of our target languages in the pre-training phase and show fine-tuning performance.
3. We evaluate the effectiveness of popular encoder-decoder and decoder-only models for multi-label emotion classification and examine the role of few-shots in improving the task.
4. We present detailed results and error analyses

[*] Equal contribution. Corr. email: tadesseit@gmail.com

[1] https://github.com/Tadesse-Destaw/EthioEmo

Proceedings of the 31st International Conference on Computational Linguistics, pages 3523–3540
January 19–24, 2025.

across languages, data sources, LLMs, and the effects of the translation test set.

2 Related Work

Emotion recognition involves identifying an underlying emotional state of individuals based on their verbal and nonverbal cues, including text, facial expressions, body language, and speech (Dadebayev et al., 2022; A.V. et al., 2024). LLMs are showing promising results for the downstream NLP tasks. Based on the training setup, the model architecture, and the use cases, LLMs can be broadly classified into encoder-only, encoder-decoder, and decoder-only types. In the past few years, there has been a significant increase in the release of decoder-only LLMs at an industry scale, and extensively used for sentiment analysis (Zhong et al., 2023; Zhang et al., 2024b). We explore emotion classification works in the following categories.

LLMs for text emotion classification: Sabour et al. (2024) proposed EmoBench to evaluate the emotional cause recognition of LLMs in English and Chinese. Liu et al. (2024) proposed EmoLLMs by fine-tuning various open-sourced LLMs for affect analysis and emotion prediction. However, these works are limited to predicting a single emotion class and lack content from languages other than English. Cageggi et al. (2023) fine-tune MT5 and evaluate FLAN and ChatGPT using few-shot prompting approaches for multi-label emotion prediction. Apart from this, the performance of other LLMs has not been assessed for multi-label emotion prediction.

Multi-label emotion classification (MLEC): To predict all possible emotions from a text, the following popular datasets were compiled: GoEmotions Demszky et al. (2020), Balanced Multi-Label Emotional Tweets (BMET) Huang et al. (2021), Romanian emotion dataset (REDv2) Ciobotaru et al. (2022), Multilingual Emotion Prediction (XLM-EMO) Bianchi et al. (2022), WASSA2023 Shared-Task 2 Ameer et al. (2023), and SemEval-2024 Task 3 (Wang et al., 2024a). Nowadays, MLEC tasks also include the corresponding intensity of each identified emotion, such as SemEval-2018 task 1 (Mohammad et al., 2018), multimodal multi-label emotion, intensity, and sentiment dialogue dataset (MEISD) (Firdaus et al., 2020), and EmoInHindi (Singh et al., 2022).

Emotion for Ethiopian languages: Emotion detection in the context of Africa generally, for Ethiopian languages specifically, has not been studied yet, except for a few sentiment analysis (negative, positive, neutral) tasks (Yimam et al., 2020; Tela et al., 2020; Muhammad et al., 2023).

Limitations of existing emotion works: Although there have been several efforts in constructing benchmark datasets and evaluations for text emotion, the existing efforts have the following shortcomings.

- The emotion research is mainly focused on English or a few other high-resource languages (Singh et al., 2022).
- Textual datasets are mostly taken from a single source, such as either news headline (Strapparava and Mihalcea, 2007), YouTube comments (Sarakit et al., 2015), Twitter (X) tweets (Mohammad et al., 2018), SMS (Ameer et al., 2023), or Facebook comments (Laabar and Zaghouani, 2024). We might not get all basic emotions from a single data source, and it is hard to generalize about emotion in this case.
- The evaluation experiments are focused on classical machine learning and deep learning approaches (Maruf et al., 2024), the current state-of-the-art LLMs' multi-label and multilingual emotional understandings are under-explored.

Towards this end, we create a multi-label EthioEmo dataset, which is constructed from various sources (news headlines, Twitter (X) posts, YouTube comments, and Facebook post comments), and each instance is annotated with one or more emotion classes. We also conduct rigorous evaluation experiments that classify emotions in multi-label settings using state-of-the-art encoder-only, encoder-decoder, and open-sourced decoder-only LLMs.

3 EthioEmo Dataset Construction

This section describes the construction of the EthioEmo dataset in detail. The driving force behind creating this dataset is the lack of an available emotion dataset in Ethiopian/African languages. Evaluating LLMs in multi-lingual and multi-label emotion understanding is another under-explored area. Moreover, emotion is language, culture, and other circumstances dependent (Sailunaz et al., 2018). EthioEmo is a new multi-label emotion dataset for four Ethiopian languages, two languages written in Ethiopic Ge'ez (`gez`) script (`amh` and `tir`) and two languages in Latin script (`orm` and som). We used Ekman's (Ekman, 1992) six basic emotion labels (anger, disgust, fear, joy, sadness, and surprise) plus neutral class.

3.1 Lexicon Collections

Lexicon entries are emotion keywords that are used to filter instances from millions of collected corpus for annotation. Based on our previous lexicon creation experiences within the Ethiopian context for sentiment analysis (Yimam et al., 2020) and hate speech (Ayele et al., 2023), we create a list of lexicon entries for each emotion class and language to ensure that each emotion class dataset is balanced and comprehensive. For example, lexicon entries of Joy emotion are "happy," "excited," and "thanks" in English. This is a step to balance the dataset by taking equal proportions from each emotion class for annotation. The Lexicon entries are adapted from an English source, NRC EmoLex Mohammad and Turney (2013), with additional manually created emotion keywords. We obtain the emotion lexicon entries in the following ways:

- Translate the English NRC EmoLex (Mohammad and Turney, 2013) lexicon into Ethiopian languages with the help of Google Translate and native speaker validations (incorrect translations are discarded).
- Collect additional emotion lexicon entries using nearest neighbors of the emotion lexicon entries from available Word2Vec and FastText word embedding models that include our target Ethiopian languages (Yimam et al., 2021; Belay et al., 2021).
- We manually add the remaining basic emotion lexicon for each language and emotion class.

We used 293 amh, 275 orm, 283 som, and 280 tir emotion query entries for six basic emotion classes. We will open-source these lexicons along with the dataset.

3.2 Data Collection

The datasets have been collected from various sources such as news portals, X/formerly Twitter, YouTube, and Facebook.

The sources are selected since they have been common data sources for previous emotion classification works and contain rich content for Ethiopian languages (Mohammad et al., 2018; Laabar and Zaghouani, 2024). These diverse sources are selected to rate the emotions that persist within texts across the sources. The statistics of the data and the sources are presented in Table 1. As part of the data preprocessing technique, language detection is applied using GeezSwitch (Gaim et al., 2022) for Ge'ez scripts and pycld3[2] for Latin scripts languages. We masked user names and URLs to prevent data privacy and confidentiality. For annotation, we select text length with a minimum of 15 characters and a maximum length of a tweet (280 characters).

Data sources	amh	orm	som	tir
Twitter (X)	2000	2700	2400	3100
Facebook	1500	600	900	600
YouTube	2000	2000	2000	2000
News headline	500	500	500	500
Total	6000	5800	5800	6200

Table 1: Data sources and sample amount taken from each source: Twitter (X) posts, Facebook post comments, YouTube video comments, and news headlines.

Regarding the period of the collected data, for Facebook, comments from posts between September and December 2023 were extracted as the data was collected at this time using the comment scraper tool[3]. For news headlines, we pulled all available BBC (https://www.bbc.com/x, where x is the name of the language) news headlines using Python script[4]. For Twitter (X), we used data scraped from 2014 to 2022 using Twitter API for academic research. For YouTube, we did not consider time span; we collected comments under the playlist/video with the most comments for the specific language using YouTube API. We applied text preprocessing such as language detection, username, URL anonymization, and over-repeated character normalization.

3.3 Data Annotation

For the data annotation, we employed native speakers for each language. Annotators were provided annotation guidelines with text examples and emotion label(s), hands-on practical training, and pilot tests before the main annotation. We compensated annotators with a payment of roughly $6 per hour on average, nearly the same as the hourly wage of Master's degree holders in Ethiopia. The detailed backgrounds of the annotators are shown in Appendix B.

We customize the POrtable Text Annotation

[2] https://pypi.org/project/pycld3/
[3] https://exportcomments.com/
[4] https://github.com/keleog/bbc_pidgin_scraper

TOol (POTATO) (Pei et al., 2022) for our in-house annotation platform. A minimum of three annotators annotated each instance. The data is annotated in multiple batches by assessing the data quality and annotators' performance, including control questions and agreements in each batch. Disagreed instances were re-annotated by new annotators, and if no agreement was reached again, the instances were excluded from the dataset. The final gold label was determined based on agreement by at least two annotators for each emotion class.

3.4 Inter-Annotator Agreement (IAA)

The most common IAA measurements, such as Cohen's kappa (Cohen, 1960), Fleiss' kappa (Fleiss, 1971), Krippendorff's alpha (Krippendorff, 2011), and bootstrapping method (Marchal et al., 2022) do not support multi-label with multiple annotators at the same time. We adopted a multi-label agreement (MLA) method proposed by Li et al. (2023) to obtain the multi-label agreement among all annotators. We also computed free marginal Randolph's Kappa scores (Randolph, 2005), a metric well-suited for measuring inter-annotator agreement in tasks involving multiple annotators.

Language	MLA	Cohen's K.	Free M.
Amharic	0.50	0.52	0.65
Afan Oromo	0.64	0.66	0.76
Somali	0.51	0.50	0.66
Tigrinya	0.53	0.57	0.68

Table 2: IAA of the EthioEmo Dataset. Multi-label Agreement (MLA) is a direct agreement between the multi-label classes and all annotators. Cohen's Kappa is a pairwise agreement between two annotators, and the result is an average pair-wise of the three annotators. Free Margin (Free M.) is calculated as a pairwise agreement between two emotion classes and takes an average.

According to the work of Sánchez-Velázquez and Sierra (2016), the IAA results in Table 2 show moderate and above agreement as Cohen's Kappa score ranges from 0.41-0.60 is moderate. For further analysis of IAA, we observe Cohen's kappa agreement for four main emotion classes (Anger, Disgust, Sadness, and Joy), and the results are Amharic: 0.74, Afan Oromo: 0.81, Somali: 0.75, and Tigrinya: 0.77, showing significantly higher scores than the scores obtained from the total of seven classes. This shows that IAA scores vary with the number of classes, as more classes generally increase the complexity of annotation, often lowering agreement scores (Stefanovitch and Piskorski, 2023). Based on Cohen's Kappa value, our IAA result is also comparable with related works, as the GoEmotion (Demszky et al., 2020) dataset IAA is 0.29 for 27 emotion classes. The highest agreement scores are reported for `Afan Oromo`. We manually go through the annotator-level data and observe that most of the annotators selected single emotions during the annotation, the reason for `Afan Oromo` having a better agreement score. This shows that the number of annotated labels by each annotator is inversely proportional to the agreement score. The overall IAA agreement score shows multi-label emotion task difficulty, a condition where an instance can have none, one, two, or all emotion classes with multiple annotators.

4 Evaluation Settings

Training and testing LLMs such as GPT-4 (OpenAI et al., 2024), Mixtral 8x22B (Jiang et al., 2024), PaLM-340B (Anil et al., 2023), and LLaMA-405B (Dubey et al., 2024) are often not feasible for academic researchers and companies with limited resources. As a result, there has been a shift towards smaller language models (Chen and Varoquaux, 2024). Our experiment includes pre-trained encoder-only, encoder-decoder, and medium-size parameter decoder-only models for scientific reproducibility. We fine-tune encoder-only models using the EthioEmo training dataset and evaluate zero-shot and in-context learning predictions with LLMs. The statistical distribution of the EthioEmo and English datasets is shown in Table 3.

Language	Train	Test	Dev	Total
Amharic	2,614	1,309	437	4,360
Afan Oromo	2,598	1,300	435	4,333
Somali	2,087	1,045	349	3,481
Tigrinya	2,865	1,435	479	4,779
English	6,327	1,232	845	8,404

Table 3: Statistics of train, test, and dev sets for EthioEmo along with SemEval-2018 Task 1 English dataset. We randomly stratify to split the EthioEmo dataset into train (60%), dev (10%), and test (30%) sets. These statistics are without the `Neutral` class as our overall experiments do not include `Neutral` class in the evaluation. Final annotated dataset statistics with no `emotion` or `Neutral` class are Amharic: 5,891, Afan Oromo: 5,690, Somali: 5,631, and Tigrinya: 6,109, a total of 23,321 instances were annotated.

4.1 Afri-centric Encoder-only Models

Considerable efforts have been dedicated to creating multilingual BERT-based encoder-only models for African languages. We select encoder-only models based on popularity, and models include at least two languages from our target languages.

We make zero-shot and fine-tuning evaluations using the following Afri-centric pre-trained language models. **AfriBERTa** (Ogueji et al., 2021) pre-trained on 11 African languages. It includes our four target languages. **AfroLM** (Dossou et al., 2022): a multilingual model pre-trained on 23 African languages, including `amh` and `orm` from Ethiopian languages. **AfroXLMR** (Adelani et al., 2024a): adaptation of XLM-R-large model (Conneau et al., 2020) (has two versions: 61 and 76 languages) for African languages including the four Ethiopian languages and high-resource languages (English, French, Chinese, and Arabic). **EthioLLM** (Tonja et al., 2024): multilingual models for five Ethiopian languages (`amh`, `gez`, `orm`, `som`, and `tir`) and English.

4.2 Open Source Decoder-only Models

From the family of decoder-only LLMs, we work with instruction-tuned versions of popular open-source models. Namely, Llama-2-7b (Touvron et al., 2023), Llama-3-8B (Meta, 2024), Llama-3.1-8B (Dubey et al., 2024), Gemma-1.1-7b (Gemma et al., 2024), and Gemma-2b (Gemma et al., 2024). From encoder-decoder, we evaluate Aya-101 (Üstün et al., 2024) — fine-tuned from mT5 (Xue et al., 2021). It is a multilingual 13B parameter model that follows instructions in 101 languages, including `amh` and `som`. We choose these models based on their popularity in the open-source community and serve as a baseline for similar NLP task evaluation. From closed-source LLMs, we include GPT-4o-mini in our evaluation as it is cost-efficient and easy to reproduce (OpenAI, 2024). We used English-based prompts for evaluating LLMs following the work by Zhang et al. (2024a); Agarwal et al. (2024) as English prompts work better than in-language prompts.

4.3 Translate Test Experiments

Following the work by Etxaniz et al. (2023), one approach to improve the performance of multilingual language models is to translate the data to English using existing machine translation systems. Our approach involves translating the EthioEmo test dataset to English to determine if English-centric models can solve the task efficiently. For the translation, we used the NLLB-200-3.3B multilingual machine translation model (Team et al., 2022).

4.4 In Context Learning (ICL)

One approach to improve the performance of LLMs is to show them examples of the task. Following the work of Zhang et al. (2024a); Agarwal et al. (2024), we use in-context learning to teach the models about the task without parameter updates by showing them input and output examples. We work with 2, 4, 6, and 8 demonstrations in our experiment and compare them with zero-shot experiments. We increase the number of contexts (k shots) by two to show the slightly increasing effects of examples. For our k-shot experiments, we applied randomly selected in-language examples from the dev set, which remained consistent across models. We used log likelihood-based evaluations using lm-evaluation-harness[5] by Gao et al. (2023) for zero-shot and few-shot LLMs experiments.

5 Results

5.1 Fine-tuned Encoder-only Models

Results of fine-tuned encoder-only models are shown in Table 4. Based on the results, AfroXLMR-76L outperforms for `amh` and `orm` with a 69.9% and 72.6% F1 score, respectively, as both languages are included in the pre-training. Examining the overall encoder-only results, AfroXLMR families perform better for the target languages. We observe that languages included in the pertaining phase perform better. Although encoder-only models demand more training data and computational resources, they still have significant room for improvement in tackling multi-label emotion classification tasks. **Pre-training is important for multi-label emotion classification task**. This is evidenced by the F1-scores we present, where the highest score is achieved by the `orm` language from AfroXLMR-76L with a score of 72.6%.

5.2 Zero-shot Experiments

We conduct a zero-shot evaluation and make the following observations, as summarized in Table 5.

Encoder-only models still have an advantage over the recently popular open-source decoder-only models for low-resource languages. We

[5] `https://github.com/EleutherAI/lm-evaluation-harness`

Model name	amh	orm	som	tir
Fine-tuned encoder-only models				
EthioLLM-small	65.3	69.4	47.1	55.7
EthioLLM-large	64.2	67.4	38.0	56.7
AfroXLMR-61L	68.3	66.5	**64.2**	**62.4**
AfroXLMR-76L	**69.9**	**72.6**	62.6	58.1
AfroLM-active-l	65.4	67.7	52.0	53.2
AfriBERTa-large	51.6	71.4	63.2	60.7

Table 4: Weighted-averaged F1-score results from fine-tuned pre-trained language models. The light-gray shows the model does not include the languages in the pre-training.

compare zero-shot results of LLMs with zero-shot and fine-tuned encoder-only models, and LLMs under-perform compared to encoder-only models. This is likely due to their initial multilingual setup of encoder-only models for low-resource languages. Cohere-aya-101 outperforms all decoder-only models with an average score of 46.12% as it is designed for multilingual and officially includes amh and som languages. Looking at the target languages, the closest performance we see between encoder-only and encoder-decoder models is in the amh language, with a score difference of 8.9% between AfroLM and Cohere-aya-101. For the encoder-only model, we can see AfroXLMR-76L takes the lead, which explains its top performance in the fine-tuning experiment. From zero-shot evaluations, Cohere-aya-101 consistently outperforms in all languages except orm. In general, the result shows how fine-tuning smaller and more efficient pre-trained language models can still outperform zero-shot performances of LLMs, which have room for improvement in multi-label emotion classification. Considerably, zero-shot or in-context learning of LLMs is not comparable with the BERT family's pre-trained model that has already seen the language in the pre-training phase. However, we compare only according to the resources that LLMs consume to fine-tune, and we expected LLMs to perform better based on their size.

5.3 Translate Test Experiments

The models struggle to classify multi-label emotions even after translating the test set to English. We conduct an experiment using the translation of the test set to investigate the reasons for poor performance in decoder-only models, as discussed in Section 4.3. Our findings reveal that even after the test set is translated into English, these models still struggle to identify emotions accurately compared to English. In particular, Cohere-aya-101 performs poorly in all EthioEmo translation test set evaluations compared to a near similar size LLaMA-3-8B-Instruct model. This might be due to either the limitations of the machine translation system employed (we do not have ground truth for further translation quality checking) or the inherent complexities of the emotion task that may not carry the same meaning across languages in the translation.

5.4 In-Context Learning Results

We do in-context learning experiments because fine-tuning LLMs can incur enormous computing costs. This approach helps improve the model's understanding ability without any parameter update.

All models benefit from two-shot examples compared to zero-shot tests. Based on the results shown in Figure 1, all our models benefit from two-shot contexts. Looking at the Ethiopian languages, we can see that they all improved their scores by showing two examples compared to zero-shot tests. However, this improvement is not shown in Gemma-1.1-7b-it, which already had good scores in the zero-shot experiment across languages. Among target languages, orm gains the highest scores in the zero-shot experiment with Gemma-1.1-7b-it model. The same pattern does not apply to som — it uses Latin script as orm, which requires further investigation. For English, Gemma-1.1-7b-it at four-shots has a better comparable result to the zero-shot.

Examining the impact of increasing the number of shots by two examples is not guaranteed to improve performance. We observed that the improvement was inconsistent and could not be guaranteed. However, there are clear performance gains from 0 to 2 shots, 2 to 6, 2 to 8, and 4 to 8 shots. This is particularly evident in all languages. The encoder-decoder Cohere-aya-101 model has a comparable best result for low-resource languages to the commercial GPT-4o-mini. **Encoder-decoder Cohere-aya-101 model outperforms the open-source LLMs**. The exception for the lowest performance for tir is that it is not included in the pre-training of the Cohere-aya-101 model. Another observation is that improvements in results are associated with the sizes of the models' size or parameters, such as from Gemma-2b-it to Gemma-1.1-7b and from LLaMA-2-7B to LLaMA-3-8B.

Pre-trained LMs	amh	orm	som	tir	eng	Average
Zero shot for encoder-only						
EthioLLM-small	31.72	12.88	30.88	32.09	37.76	29.07
EthioLLM-large	14.38	32.35	10.53	10.94	37.87	21.21
AfroXLMR-61L	22.05	39.68	21.12	22.30	43.00	29.63
AfroXLMR-76L	28.62	35.81	14.86	15.94	24.38	23.92
AfroLM-active-l	39.91	25.60	15.92	**35.63**	33.93	30.20
AfriBERTa-large	25.67	15.57	19.98	35.39	26.38	24.60
Zero shot for decoder-only						
Gemma-2b-it	10.22	7.81	14.26	7.87	46.4	17.37
Gemma-1.1-7b-it	27.94	34.87	25.87	19.55	65.73	34.79
LLaMA-2-7b-chat-hf	17.35	19.24	22.05	12.97	54.07	25.14
LLaMA-3-8B-Instruct	28.18	26.91	29.29	19.73	66.74	34.17
Llama-3.1-8B-Instruct	20.58	24.10	22.07	10.28	51.17	25.25
Cohere-aya-101	48.80	33.65	43.00	38.97	66.20	46.12
Zero shot for closed models						
GPT-4o-mini	53.86	47.84	52.04	35.47	70.98	52.04
Zero shot for decoder-only translated to English						
Gemma-2b-it	30.91	28.66	31.43	22.54		28.39
Gemma-1.1-7b-it	45.05	47.86	44.72	35.35		43.25
LLaMA-2-7b-chat-hf	35.48	34.27	34.58	25.19		32.38
LLaMA-3-8B-Instruct	48.26	48.55	46.46	40.93		46.06
Llama-3.1-8B-Instruct	28.59	34.43	32.28	21.66		29.24
Cohere-aya-101	44.95	43.39	41.52	31.63		40.37
Translated zero shot for closed models						
GPT-4o-mini	55.89	51.59	51.14	47.60		51.56

Table 5: Zero-shot experiment results from encoder-only and decoder-only models (weighted-averaged F1-score) across languages. The translated test is by translating EthioEmo test set to English. The light-gray background indicates AfroLM does not include the languages in the pre-training.

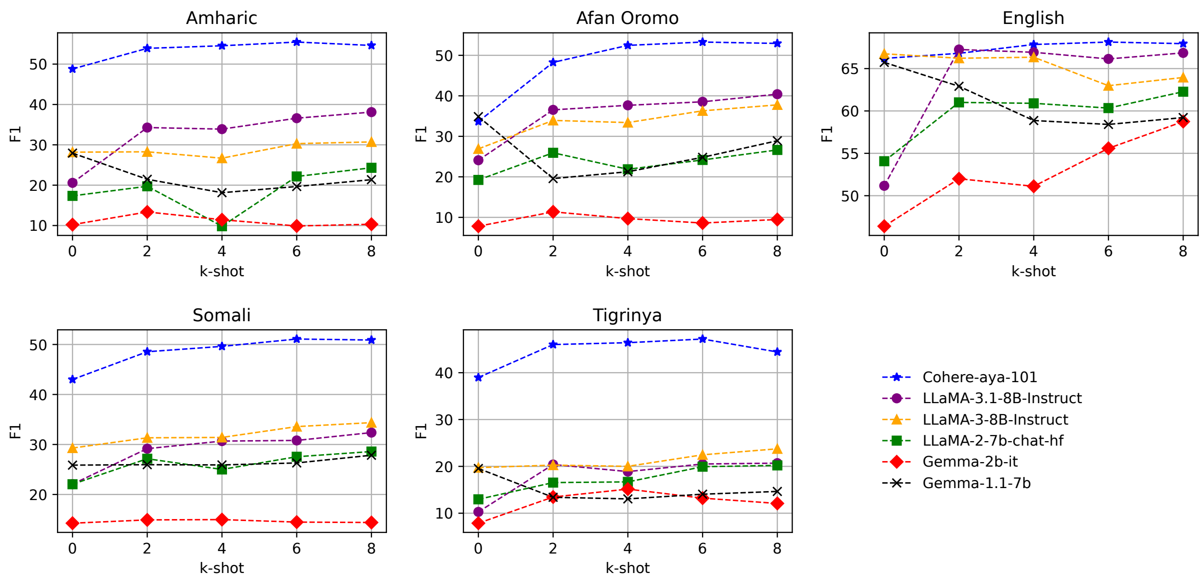

Figure 1: In-context learning (ICL) experiments with k-shots and languages.

5.5 Prompt Sensitivity Experiment

Role-based prompt gained more results from LLMs. A drawback of the prompting evaluation is the model sensitivity to prompts, where slight changes in instruction can lead to large differences in performance (Sun et al., 2023). To handle this, we use the following three prompts: (1) **generic**: a prompt which does not give information about the task, used in (Liu et al., 2024); (2) **task-based**: describes the given task (Edwards and Camacho-Collados, 2024); (3) **role-based**: a new prompt which gives more information, including `"You are a helpful AI assistant that can identify emotions from text"`. All prompting results presented in this paper are averages of the three prompts. For reproducibility of the experiment, the prompts are shown in Appendix 2, and the results of each prompt are presented in Appendix G.3.

Generic prompt :

Identify all applicable emotions for the given text from (anger, disgust, fear, sadness, joy, surprise).
Text: """{text}"""
Answer:

Task-based prompt :

Categorize the text's emotional tone as the presence of one or more of the given emotions (anger, disgust, fear, sadness, joy, surprise) only. No explanation is needed.
Text: """{text}"""
Answer:

Role-based prompt :

You are a helpful AI assistant that can identify emotions from text. Categorize the text's emotional expression, classifying it as one or more of the specified emotions ['anger', 'disgust', 'fear', 'sadness', 'joy', 'surprise'] that reflect the writer's state of mind. No explanation is needed.
Text: """{text}"""
Answer:

Figure 2: The three prompts used for decoder-only zero-shot and in-context learning experiments

6 Error Analysis and Discussion

Task difficulty: Our analysis shows that the task is not easily solvable by any of the methods. This shows the significance of this task in evaluating the existing models and observing that multi-label emotion classification needs more exploration, even for high-resource languages such as English. Some of the difficulties include 1) inability to know the exact feelings of the writers in sarcastic texts — needs context, and 2) ambiguity between some emotion classes such as `Anger` and `Disgust` (for example, in 61 instance annotations, `Anger` and `Disgust` appear together from 85 disagreed `amh` instances). The fact that the task is difficult also means a great deal of work to advance further research in emotion detection and analysis, which shows that EthioEmo dataset is a useful contribution to evaluating the upcoming advanced models.

Data sources from news headlines mostly exhibit none of the six basic emotion classes, while other sources are better for the basic emotions. We visualize the statistics of emotion distributions across languages and data sources are shown in Appendix F; instances sourced from news headlines are almost `Neutral - do not have any of the basic emotions`. Emotion classes such as `Anger`, `Disgust`, and `Joy` are shared on Facebook comments and Twitter (X) posts. YouTube comments include all basic emotions — a better source for Ethiopian languages' emotion data that has the rare `Surprise` emotion class. The statistics of emotion distributions across languages and data sources are shown in Appendix F.

Challenges in emotion annotation : Obtaining consistent annotations for an NLP dataset, especially for emotion, is challenging. This is due to several reasons, including 1) difficulty in knowing the exact feeling of the writers in sarcastic texts, 2) differences in human experience that impact how they perceive emotion in text, 3) sometimes annotators depend only on emotion keywords present in the text during annotation, 4) ambiguity between some emotion classes such as *Anger* and *Disgust*, and 5) reports of third person sayings as the writer's emotion are some of the challenges encountered during the annotation.

Experiment error analysis: Regarding the dataset emotion distribution, the dataset has more `Anger` and `Disgust`. Firstly, this is common also in other emotion datasets (Mohammad et al., 2018; Demszky et al., 2020; Wang et al., 2024a). Secondly, one reason is because of the conflict situations (in the year 2023) in some parts of Ethiopia and in the global context, such as the Hamas-Israel and Russia-Ukraine wars.

We go through the predicted test file from the best-performing encoder-only model, AfroXLMR-76L, with the help of experts from each language. The following are the most common cases of the incorrect prediction of emotions. 1) While the gold labels of an instance have more than one emotion, the models predict a single emotion label and vice versa — a common issue in a multi-label classification problem. 2) The model classifies based on some emotion keyword/emojis in the text, not the whole context. **This shows that text with emojis is straightforward for the models to predict emotions** and is aligned with the work of Liegl and Furtner (2024). 3) The model fails due to incomplete text and grammatical errors. This is mainly due to the limited length of tweets and the informal writing style of social media (Belay et al., 2022). 4) The model fails to categorize a text into specific emotion classes, resulting in nothing predicted. This problem is shown in encoder-decoder models. In multi-label classification, encoder-only works in OneVsRest (One vs other) approach (Goštautaitė and Sakalauskas, 2022) which is predicting whether each emotion is present or not separately, for instance, `Anger` or `Not anger`, `Fear` or `Not Fear`. When the model responds `Not` for all emotions, the result will be nothing predicted. On the other hand, decoder-only models such as GPT-4o-mini (OpenAI et al., 2024) show an over-predicted problem, assigning more emotion classes while most of the instances have a single emotion class. For these above-mentioned cases, emotion instance examples are shown in Appendix E for the corresponding language and case number.

7 Conclusion and Future Work

In light of the growing interest in creating challenging NLP tasks to assess the abilities of LLMs, language- and culture-specific datasets are becoming crucial (Wang et al., 2024c; Adelani et al., 2024b). This work presented a multi-label emotion dataset (EthioEmo) and an evaluation of multi-label emotional understanding of encoder-only, encoder-decoder, and decoder-only language models. The dataset provides diversity regarding the data source (X/Twitter posts, YouTube comments, Facebook comments, and news headlines) and four Ethiopian languages with available English dataset for evaluation). We reported strong baseline results using various experimental settings such as fine-tuning encoder-only models, translated test sets, prompt sensitivity, zero-shot, and impacts of increasing the number of shots for in-context learning evaluations. Encoder-only Afri-centric models that include target languages during the pre-training phase are the best for the classifications of the EthioEmo dataset. In general, the results show that fine-tuning encoder-only language models can still outperform the few-shot approaches of LLMs. The open-source Cohere-aya-101 model outperformed other LLMs next to the commercial GPT-4o-mini. This paper focused on evaluating state-of-the-art open-source LLMs with the least parameters for scientific reproducibility. Fine-tuning open-source and evaluating closed-source LLMs are out of the scope of this work and are the next works. We believe this dataset and results can be employed as a baseline in the future for better multi-label emotion classification tasks. Resources such as lexicons, annotation guidelines, and datasets are publicly available for further investigation.

Limitations

In this work, we present and evaluate the EthioEmo dataset using Afri-centric pre-trained language models and open-source LLMs. Despite our efforts, the following are limitations of this work.

Imbalanced data: Even if it is impossible to balance emotion data, we tried to balance the emotions using lexicon entries. One of the limitations of this work is that the distribution of the emotion classes is imbalanced. Having more balanced data would be better. However, the nature of the task itself makes it challenging to balance each emotion class because all emotions are not expressed equally in the data source platforms.

Translation effect on emotions: To evaluate generative models, we translate the EthioEmo test dataset to English to know if the prediction difficulties come from the task's nature or language understanding. However, this translation will have quality and context effects on the emotion itself as emotions are culture and language-dependent.

Acknowledgments

This work was carried out as a part of the Afri-Hate project of data annotation with the support of the Lacuna Fund, an initiative co-founded by The Rockefeller Foundation, Google.org, and Canada's International Development Research Center. The authors thank the CONAHCYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award. We thank OpenAI for providing API credits to Masakhane. We also acknowledge the support of the LT Group, the University of Hamburg, for hosting the annotation tools.

References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024a. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 226–245, St. Julian's, Malta.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, and Pontus Stenetorp. 2024b. Irokobench: A new benchmark for african languages in the age of large language models.

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, MN, USA.

Iqra Ameer, Noman Ashraf, Grigori Sidorov, and Helena Gómez Adorno. 2020. Multi-label emotion classification using content-based features in twitter. *Computación y Sistemas*, 24(3):1159–1164.

Iqra Ameer, Necva Bölücü, Hua Xu, and Ali Al Bataineh. 2023. Findings of WASSA 2023 shared task: Multi-label and multi-class emotion classification on code-mixed text messages. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 587–595, Toronto, Canada.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, and others. 2023. PaLM 2 technical report. *Preprint*, arXiv:2305.10403.

Geetha A.V., Mala T., Priyanka D., and Uma E. 2024. Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *Information Fusion*, 105:102218.

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023. Exploring Amharic hate speech data collection and classification approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 49–59, Varna, Bulgaria.

Tadesse Destaw Belay, Abinew Ali Ayele, Getie Gelaye, Seid Muhie Yimam, and Chris Biemann. 2021. Impacts of homophone normalization on semantic models for Amharic. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 101–106, Bahir Dar, Ethiopia.

Tadesse Destaw Belay, Seid Muhie Yimam, Abinew Ayele, and Chris Biemann. 2022. Question answering classification for Amharic social media community based questions. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 137–145, Marseille, France.

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

Gioele Cageggi, Emanuele Di Rosa, and Asia Uboldi. 2023. App2check at emit: Large language models for multilabel emotion classification. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop*, Parma, Italy.

Lihu Chen and Gaël Varoquaux. 2024. What is the role of small models in the LLM era: A survey. *Preprint*, arXiv:2409.06857.

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.

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.

Didar Dadebayev, Wei Wei Goh, and Ee Xion Tan. 2022. Eeg-based emotion recognition: Review of commercial eeg devices and machine learning techniques. *Journal of King Saud University - Computer and Information Sciences*, 34(7):4385–4401.

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.

Jiawen Deng and Fuji Ren. 2020. Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning. *IEEE Transactions on Affective Computing*, 14(1):475–486.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA.

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, UAE.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. Ethnologue: Languages of the World. Twenty-third edition. Dallas, Texas: SIL International. `http://www.ethnologue.com/`. [Accessed 01-06-2024].

Aleksandra Edwards and Jose Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10058–10072, Torino, Italia.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english? *arXiv preprint arXiv:2308.01223*.

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.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Fitsum Gaim, Wonsuk Yang, and Jong C. Park. 2022. GeezSwitch: Language identification in typologically related low-resourced East African languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6578–6584, Marseille, France.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation. [Accessed July-07-2024].

Team Gemma, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Daiva Goštautaitė and Leonidas Sakalauskas. 2022. Multi-label classification and explanation methods for students' learning style prediction and interpretation. *Applied Sciences*, 12(11).

Chenyang Huang, Amine Trabelsi, Xuebin Qin, Nawshad Farruque, Lili Mou, and Osmar Zaïane. 2021. Seq2Emo: A sequence to multi-label emotion classification model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4717–4724, Online.

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.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2022. A review on text-based emotion detection–techniques, applications, datasets, and future directions. *arXiv preprint arXiv:2205.03235*.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2023. A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection. *Artificial Intelligence Review*, 56(12):15129–15215.

Sanaa Laabar and Wajdi Zaghouani. 2024. Multi-dimensional insights: Annotated dataset of stance, sentiment, and emotion in Facebook comments on Tunisia's July 25 measures. In *Proceedings of the Second Workshop on Natural Language Processing for Political Sciences @ LREC-COLING 2024*, pages 22–32, Torino, Italia.

Sheng Li, Rong Yan, Qing Wang, Juru Zeng, Xun Zhu, Yueke Liu, and Henghua Li. 2023. Annotation quality measurement in multi-label annotations. In *International Conference on Natural Language Processing and Chinese Computing*, pages 30–42, Cham, Switzerland.

Simon Liegl and Marco R Furtner. 2024. Emotional leader communication in the digital age: An experimental investigation on the role of emoji. *Computers in Human Behavior*, 154:108148.

Xuan Liu, Tianyi Shi, Guohui Zhou, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. 2023. Emotion classification for short texts: an improved multi-label method. *Humanities and Social Sciences Communications*, 10(1):1–9.

Zhiwei Liu, Kailai Yang, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. *Preprint*, arXiv:2401.08508.

Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3659–3668, Gyeongju, Republic of Korea.

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.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date — ai.meta.com. `https://ai.meta.com/blog/meta-llama-3/`. [Accessed 01-06-2024].

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.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Shamsuddeen Muhammad, Idris Abdulmumin, Abinew Ayele, Nedjma Ousidhoum, David Adelani, Seid Yimam, Ibrahim Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Alipio Jorge, Pavel Brazdil, Felermino Ali, Davis David, Salomey Osei, Bello Shehu-Bello, Falalu Lawan, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Destaw Belay, Wendimu Messelle, Hailu Balcha, Sisay Chala, Hagos Gebremichael, Bernard Opoku, and Stephen Arthur. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore.

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.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. `https://openai.com/`. [Accessed Augest-07-2024].

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

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.

Justus J. Randolph. 2005. Free-marginal multirater kappa (multirater k[free]): An alternative to fleiss' fixed-marginal multirater kappa.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Juanzi Li, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models.

Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. 2018. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):28.

Octavio Sánchez-Velázquez and Gerardo Sierra. 2016. Let's agree to disagree: Measuring agreement between annotators for opinion mining task. *Research in Computing Science*, 110:9–19.

Phakhawat Sarakit, Thanaruk Theeramunkong, Choochart Haruechaiyasak, and Manabu Okumura. 2015. Classifying emotion in thai youtube comments. In *2015 6th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*, pages 1–5.

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.

Nicolas Stefanovitch and Jakub Piskorski. 2023. Holistic inter-annotator agreement and corpus coherence estimation in a large-scale multilingual annotation campaign. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 71–86, Singapore.

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

Jiuding Sun, Chantal Shaib, and Byron C Wallace. 2023. Evaluating the zero-shot robustness of instruction-tuned language models. *arXiv preprint arXiv:2306.11270*.

Jie Tao and Xing Fang. 2020. Toward multi-label sentiment analysis: a transfer learning based approach. *Journal of Big Data*, 7(1):1–26.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, and others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Abrhalei Tela, Abraham Woubie, and Ville Hautamaki. 2020. Transferring monolingual model to low-resource language: the case of tigrinya. *arXiv preprint arXiv:2006.07698*.

Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, and others. 2024. EthioLLM: Multilingual large language models for Ethiopian languages with task evaluation. *arXiv preprint arXiv:2403.13737*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024a. SemEval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2039–2050, Mexico City, Mexico.

Kaipeng Wang, Zhi Jing, Yongye Su, and Yikun Han. 2024b. Large language models on fine-grained emotion detection dataset with data augmentation and transfer learning. *Preprint*, arXiv:2403.06108.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and others. 2024c. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. 2020. Exploring Amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1048–1060, Barcelona, Spain.

Seid Muhie Yimam, Abinew Ali Ayele, Gopalakrishnan Venkatesh, Ibrahim Gashaw, and Chris Biemann. 2021. Introducing various semantic models for amharic: Experimentation and evaluation with multiple tasks and datasets. *Future Internet*, 13(11).

Miaoran Zhang, Vagrant Gautam, Mingyang Wang, Jesujoba O Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach. 2024a. The impact of demonstrations on multilingual in-context learning: A multidimensional analysis. *arXiv preprint arXiv:2402.12976*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024b. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A EthioEmo Languages

There are more than 2000 languages spoken in the African continent, and more than 80 of them are spoken in Ethiopia[6]. Amharic, Afan Oromo, Somali, and Tigrinya are the top four languages in Ethiopia by the number of speakers.

Amharic (amh): is a Semitic language written in Ge'ez script, known as Fidel, which consists of 33 primary characters, each with seven vowel sequences. It is the second most widely spoken Semitic language, next to Arabic.

Afan Oromo (orm): is an Afro-Asiatic language written in Latin script. It is the most widely spoken language in Ethiopia and the third most widely spoken in Africa, next to the Arabic and Hausa languages. It is mostly spoken in the Horn of Africa, including Ethiopia, Kenya, and Somalia alone.

Somali (som): is an Afro-Asiatic language belonging to the Cushitic group. It is spoken in Ethiopia, Somaliland, Kenya, and Somalia. It is the third most widely spoken language in Ethiopia.

Tigrinya (tir): is a Semitic language spoken in the Tigray region of Ethiopia and Eritrea. The language uses Ge'ez script with additional Tigrinya alphabets and is closely related to Ge'ez and Amharic (Eberhard et al., 2024).

B Annotators Background

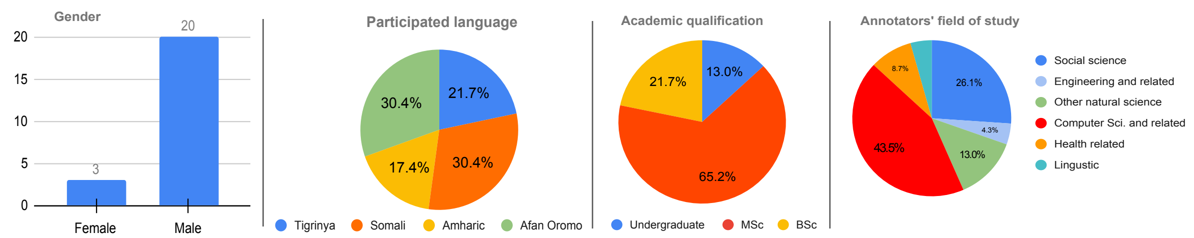

Figure 3: Backgrounds of Annotators: gender, language participated, academic qualification, and field of study.

C Number of Emotion Labels Per Instance

Figure 4 shows the number of emotion label(s) distribution across languages. As we can see, most of the dataset for all languages has a single emotion class. Of the `amh` dataset, 88% has a single label, 11.7% has two emotion labels, and 0.17% has three labels. In the `orm`, 92.9% has a single label, 6.9% two labels, and 0.17% three labels. In `som`, 94.8% has single labels, 5.7% two labels, and only three instances have three labels. In `tir`, 82.3% has single labels, 12.4% two labels, and 0.36% three labels. EthioEmo dataset is also used for multiclass emotion classification for future work as instances with more than two labels are less.

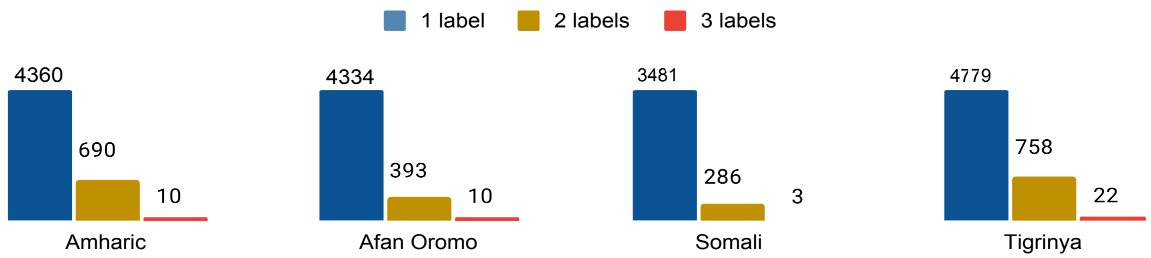

Figure 4: Number of emotion labels per instance across languages

[6]https://www.statista.com/statistics/1280625/number-of-living-languages-in-africa-by-country/

D Experiment Hyper-parameters

We make fine-tune encoder-only models using FLAIR framework (Akbik et al., 2019) with the following hyper-parameters: model_max_length = 512 (except AfroLM model_max_length is 256 as the models built it up), learning_rate = 5.0e-5, mini_batch_size = 8, and max_epochs=3, as recommended in the BERT paper (Devlin et al., 2019). We test decoder-only models with temperature = 0 and batch_size = 1.

E Emotion Examples for Error Analysis

#	Text	Gold	Pred.
1	*amh*: ጥንካሬሽ የሚገረም ነው በእባ ጨረስኩት የሰው ጭካኔ ግን 😭	sadness, surprise	sadness
	orm: Humni shiraan guddate shiraan jiraachuu filata . Haata'uutii shirri qaama biraa irratti hojjat balleessa . Warri ODP amala badaa.	disgust	anger, disgust
	tig: ኣንታ እወ ጭንቀት ሞይትና እናነበርና ናብ ዘይምንባር ተፀጊዕና ምስ ህዝብካ ኮይንካስ ኩሉ ትርኢ ሓደ ዓመት ሓሊፉና ኣሎ ትግራይ ትስዕር *tir*: ከምዚኦም ዝኣመሰሉ ደመኛታት ጸላእትና ኩላትና ተኣኪብና ብዓውታ ብዘይፍርሒ እንተዘይመኪትናዮምን ሓድነትና እንተዘይርኢናዮምን ሃገርና	fear, sadness	sadness, anger
	som: USER Madooobow adeeer waaarimeyside war hakaaa haro magaaaladiii aaad dhaqaaalaheeeda intaaasooo sanadoood dhuuuqeysay iney sidaaas u burbursanaaato waaa kugu ceeeb dhowr laaammmi lama rabeee inta jid eee magaaalada ka baxda ilaaa iyo kuwa Dhooobley iyo Afmadow tago ayaaa lagaaa rabay inaaad laaammmi saaarto	disgust	anger, disgust
2	*amh:* የሰማይ የምድር ጌታ በቤታቸው በኑሮዋቸው በልጆቻቸው በዘመድ ባዝማዳቸው ችግር አዘን መከራ **አዘን** ይግባባቸው	disgust	sadness
	orm: Biyyyaaa Sanaaa keeesaaa seeeriii hin jiruuu kani kan godhuuu abbbiyyyiiidha Ummmani Oromooo ganaaa fixxxaaa Dinnnaaa fanaaa hiriruuuraaa otttooo tokkkooo tannneee dinnnaaa ofiraaa fincileee **gaaariiidha**	anger	joy, anger
	orm: Takele uma motumaan kun hin kufaa shakiin hin jiruu **wayyanenuu** hin kufttee sodaa lubbutiif jetee umataa kee baldhaa hin ganiin dhugaa irraa dhabadhuu dinaa wajiin harkaa hin dhahiin URL	anger	disgust
	tir: እስኪ ሓድሽ ድሃይ ሃብና እዚ ምድግጋም እንታይ ዋጋ ኣሎዎ አረ ብጭንቀት **ክንመውት** እና በጃኹም ድሃይ ሃቡና	fear	sadness
3	*amh*: ትክክል ነው የዘነጉት ግን ይህ ሰው በስልጣን ከቀጠለ አገርም ሊኖራቸው እንደማይችል ነው። የ USER ስጋት ለአማራ ብቻ ነው የሚለው ስህ	fear	disgust, anger
	tir: ግፍዒ ሰቃይ ስእነት ዋላከዋ ዘይደለ አንተ ኾነ ኣብ ? ጠራይ ከም ዝርከብ ገርካ ምግራምካ ገሪሙኒ ምናልባት ስደት	surprise	sadness

Table 6: Examples of emotion texts for cases discussed in the experiment error analysis, Section 6; `amh`, `orm`, `som`, and `tir` are the languages used in the example.

F Data Source and Emotion Distribution

In this section, we visualize the EthioEmo dataset emotion distributions across data sources: Twitter (X), YouTube, Facebook, and news headlines and languages: `amh`, `orm`, `som`, and `tir`. Figure 5 shows general emotion distribution across data sources for the EthioEmo dataset. Figure 6 shows emotion distribution across languages. Figures 7, 8, 9, and 10 show emotion distribution across languages and data sources: Facebook comments, YouTube comments, Twitter (X), and news headlines, respectively. Figure 10 shows that the news headlines almost do not have any of the basic emotions.

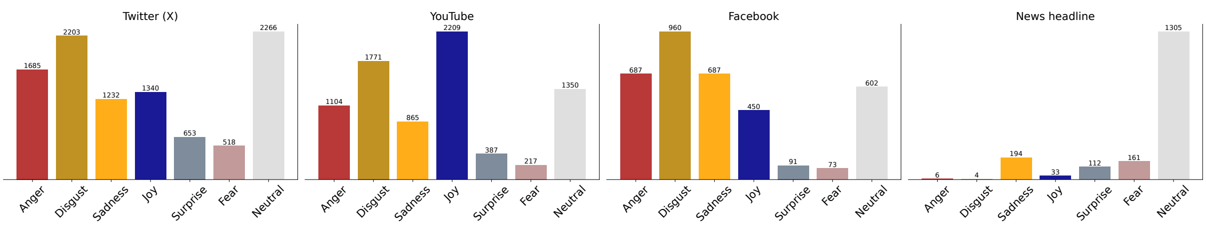

Figure 5: Emotion distribution in the data sources for EthioEmo dataset

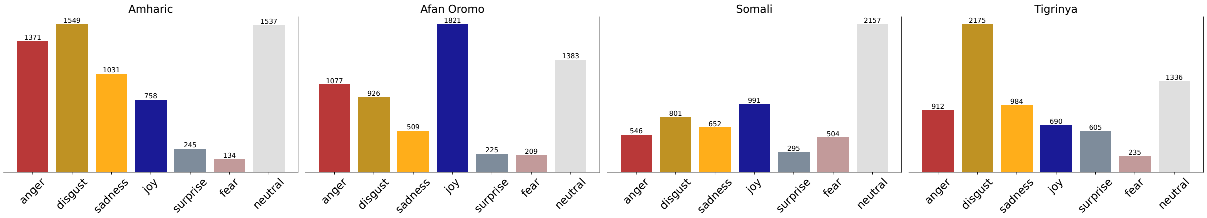

Figure 6: Emotion distribution for EthioEmo dataset across languages

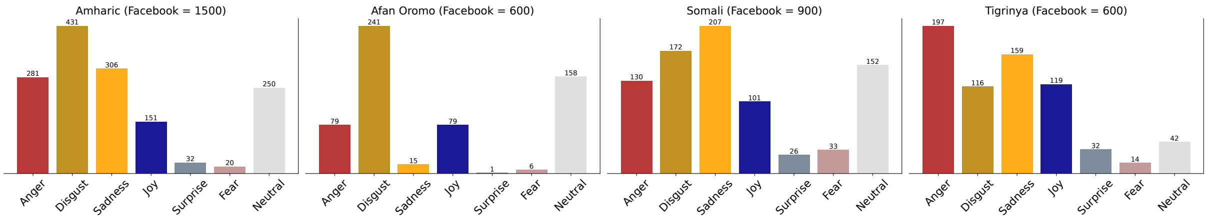

Figure 7: Facebook comment emotions distribution across languages from the given quota

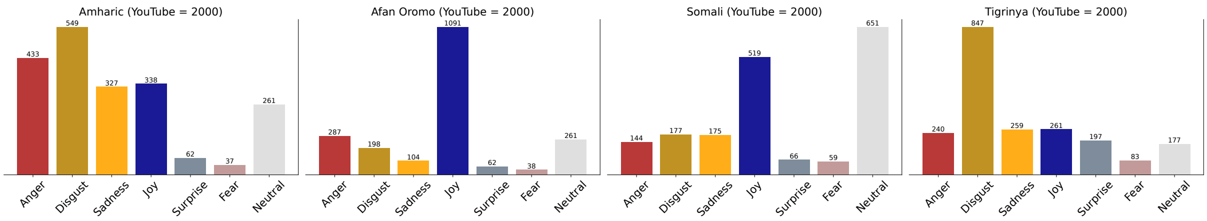

Figure 8: YouTube comment emotions distribution across languages from the given quota

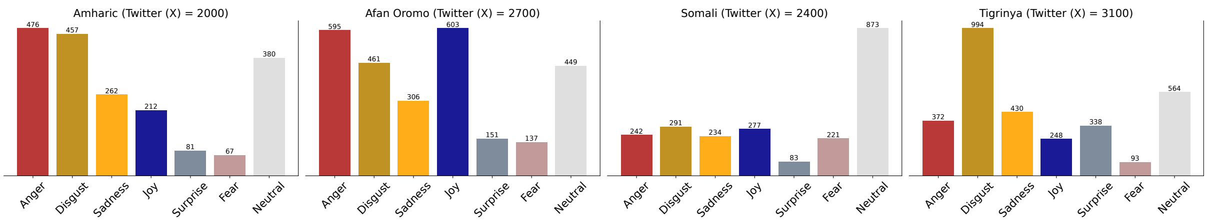

Figure 9: Twitter (X) post emotions distribution across languages from the given quota

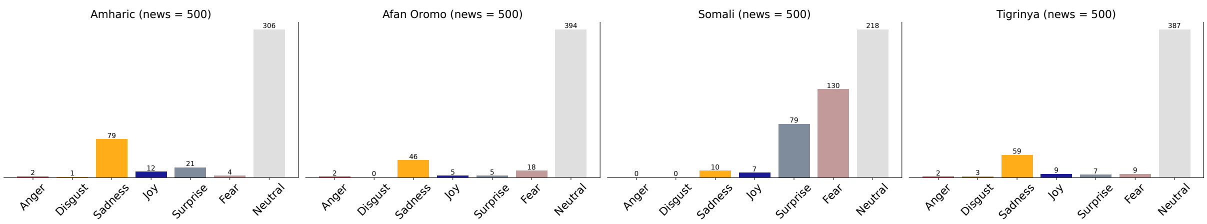

Figure 10: News headlines emotions distribution across languages from the given quota

G Additional Results

G.1 Encoder-only Experimental Results

We compared our experimental results using a weighted-averaged F1-score. Table 7 shows additional multi-label evaluation metrics such as multi-label accuracy, macro F1-score, and micro F1-score across languages and encode-only models.

	Amharic (amh)			**Afan Oromo** (orm)			**Somali** (som)			**Tigrinya** (tir)		
Pre-trained LMs	**Acc**	**Mac F1**	**Mic F1**	**Acc**	**Mac F1**	**Mic F1**	**Acc**	**Mac F1**	**Mic F1**	**Acc**	**Mac F1**	**Mic F1**
EthioLLM-s-70K	50.3	58.8	69.0	60.94	61.4	70.0	40.8	42.4	51.0	48.0	46.3	59.8
EthioLLM-l-70K	48.6	53.9	65.1	60.0	55.7	68.9	31.2	32.2	43.6	48.9	46.8	60.3
Afro-xlmr-large-61L	54.0	68.3	68.4	58.4	55.8	68.0	53.4	61.7	64.9	51.4	55.4	64.5
Afro-xlmr-large-76L	55.6	67.5	70.0	63.3	66.3	72.9	53.6	59.4	64.3	49.1	48.4	61.4
AfroLM-active-learning	50.0	60.7	65.6	59.1	52.2	66.2	40.9	48.2	53.2	36.6	33.8	49.5
Afriberta-large	67.5	64.2	67.7	62.4	63.7	71.8	53.7	60.1	64.5	49.3	52.8	62.6

Table 7: Additional results of encoder-only models. Acc - multi-label average accuracy, Mac F1 - macro F1-score, and Mic F1 - micro F1-score.

G.2 Emotion Class-based Results

Table 8 shows class-level emotion results from the three encoder-only models. We discovered that the emotion class with less dataset distribution performs low, for example, `Fear` class in `orm` and `tir` languages. In overall performance across languages, `som` and `tir` have low performance; this might be because of the amount of corpus in the pre-training and the data in the emotion classes.

	EthioLLM-s-70K				**Afro-xlmr-large-61L**				**Afro-xlmr-large-76L**				**AfriBERTa-large**			
Emotions	**amh**	**orm**	**som**	**tir**	**amh**	**orm**	**som**	**tir**	**amh**	**orm**	**som**	**tir**	**amh**	**orm**	**som**	**tir**
Anger	56.3	55.1	10.0	19.5	59.6	55.1	41.1	25.6	58.7	61.2	30.2	20.2	58.3	57.4	39.7	29.2
Disgust	68.4	68.9	36.7	71.9	66.7	629	58.3	77.2	71.5	68.2	57.4	73.6	68.6	68.1	53.8	74.6
Fear	21.7	30.2	57.8	2.8	48.7	12.3	69.1	27.1	53.9	42.0	70.7	00.0	40.7	34.9	73.2	18.2
Sadness	74.7	55.2	53.7	53.9	75.0	54.1	71.4	62.9	77.8	63.1	68.7	58.4	74.7	62.0	70.4	61.7
Joy	71.2	85.6	72.4	56.9	82.3	84.2	77.5	63.2	81.2	87.0	79.8	64.2	76.9	87.5	79.0	60.9
Surprise	60.6	73.0	23.9	73.0	65.5	66.7	52.9	76.5	62.3	76.0	49.6	74.1	66.2	72.4	44.6	72.3

Table 8: Class-based emotion F1 results from selected fine-tuned encoder-only models

G.3 Results Across Prompts, Languages, and k-shots

The details of the prompts are shown in Figure 2. Prompt 1 is a generic prompt, Prompt 2 is a task-based prompt, and Prompt 3 is a role-based prompt.

		Gemma-2b-it					Gemma-1.1-7b					LLaMA-2-7b-chat-hf				
	Lang	0	2	4	6	8	0	2	4	6	8	0	2	4	6	8
Prompt 1	amh	8.33	16.98	12.98	11.14	12.99	31.55	20.88	17.11	19.03	20	3.59	49.74	49.68	21.74	25.62
	eng	47.27	59.92	58.6	58.91	61	67.5	64.45	58	57.51	58.72	49.09	60.36	59.82	58.01	59.63
	orm	7.05	18.04	11.57	10.57	11.01	31.92	21.4	19.79	24.21	28.39	11.44	25.86	20.24	22.26	24.6
	som	14.16	16.12	14.88	15.47	14.81	20.8	25.83	25.73	26.21	27.57	13.92	28.13	25.26	29.14	29.38
	tir	7.82	18.48	16.98	15.59	15.52	27.84	13.56	12.86	12.71	14.04	2.5	16.42	16.17	20.21	20.91
Prompt 2	amh	11.57	10.21	9.78	8.2	8.36	25.61	21.26	18.42	19.18	21.27	23.83	18.87	19.52	21.17	21.53
	eng	18.82	48.21	47.69	52.49	57.36	65	60.54	58.64	58.58	58.63	54.16	61.52	61.49	60.56	62.98
	orm	6.99	6.64	7.77	7.55	8.17	36.36	18.54	21.11	25.19	29.4	23.99	25.33	21.23	23.08	27.47
	som	15.13	12.31	13.11	12.51	12.36	28.73	25.07	25.57	26.61	27.41	26.41	27.75	25.15	27.68	28.37
	tir	8.73	9.68	15.91	13.32	10.52	15.6	12.55	12.93	13.42	14.79	22.07	15.92	15.23	19.38	19.06
Prompt 3	amh	10.76	12.9	11.57	10.34	9.6	26.65	22.23	18.88	20.8	22.8	24.64	20.59	20.29	23.58	25.79
	eng	43.11	47.9	47.05	55.36	57.89	66.49	63.74	59.98	59.14	60.31	59.97	61.15	61.39	62.46	64.21
	orm	9.39	9.39	9.74	7.68	9.23	36.32	18.76	22.89	25.07	28.95	22.28	26.62	24.14	27.17	27.85
	som	13.5	16.32	16.94	15.45	16	28.09	26.99	26.42	26.11	28.56	25.82	25.57	24.63	25.82	27.12
	tir	7.06	12.22	12.55	10.75	10.22	15.2	14.09	13.39	16.04	15.11	14.34	17.29	18.67	20.2	20.64

		LLaMA-3-8B-Instruct					LLaMA-3.1-8B-Instruct					CohereForAI__aya-101				
Prompt 1	amh	15.38	29.22	27.92	31.51	32.79	21.63	34.25	33.58	37.96	39.52	49.98	54.7	55	56.59	57.54
	eng	64.82	66.22	64.82	63.54	64.41	54.64	68.21	67.42	66.4	67.22	63.88	65.97	67.42	67.93	67.16
	orm	25.9	33.43	33.77	37.65	39.12	29.71	36.24	37.75	38.25	40.67	31.71	48.8	53.47	53.73	54.75
	som	26.01	32.65	33.49	34.77	34.46	23.19	34.99	38.55	37.63	40.31	36.49	47.39	49.18	50.71	51.84
	tir	8.4	21.72	19.64	22.37	23.45	8.4	20.52	19.25	21.41	22.64	41.52	48.96	50.09	50.84	50.01
Prompt 2	amh	36.95	27.63	27.25	30.15	30.55	32.73	33.9	34.24	35.65	37.78	47.38	53.73	54.96	56.12	66.66
	eng	67.12	65.87	63.55	61.94	63.58	65.46	66.47	66.49	66.57	67.39	65.21	65.33	65.83	66.75	66.66
	orm	24.6	33.29	31.72	35.72	37.24	31.83	37.24	37.9	39.36	41.25	35.42	47.82	51.52	53.9	54.27
	som	30.22	31.63	31.34	34.55	34.89	29.57	18.08	16.69	17.42	18.45	45.32	49.35	49.06	51.66	51.33
	tir	31.32	17.72	18.88	19.86	22.39	15.34	18.08	16.69	17.42	18.45	38.58	47.86	49.04	50.67	49.37
Prompt 3	amh	32.22	27.95	24.9	28.44	28.83	7.38	34.68	33.86	36.2	37.04	49.03	53.29	53.7	53.7	50.58
	eng	68.27	66.57	64.66	63.45	63.85	33.4	67.07	66.84	65.46	65.93	69.52	69.11	70.26	69.71	70
	orm	30.24	35.06	34.71	35.52	36.94	10.77	36.13	37.32	37.97	39.32	33.82	48.15	52.4	52.21	49.8
	som	31.64	29.73	29.39	31.41	33.77	13.44	34.45	36.81	37.38	38.38	17.18	18.9	50.62	50.91	49.51
	tir	19.46	21.37	21.51	25.17	25.41	7.11	22.59	20.72	22.69	20.98	36.81	41.2	40.1	40.05	33.92

Test-set translated results

		Gemma-2b-it					Gemma-1.1-7b					LLaMA-2-7b-chat-hf				
Prompt 1	amh	33.29	36.1	35.57	37.83	40.6	42.67	43.12	39.06	40.31	41.53	24.7	43.7	41.27	61.65	46.01
	orm	29.7	34.7	40.41	41.74	44.54	46.52	40.71	39.7	42.48	41.9	22.64	43.79	41.16	44.45	44.71
	som	29.24	38.42	40.27	40.74	42.85	42.71	47.29	45.73	45.26	46.01	28.44	46.24	45.71	46.03	48.89
	tir	25.57	26.5	29.07	30.47	30.22	33.43	32	31.6	31.65	33.11	19.88	31.09	31.02	31.15	37.08
Prompt 2	amh	33.77	31.67	32.71	35.64	37.45	45.44	40.4	40.73	40.58	42.45	41.92	43.39	41.75	43.43	45.51
	orm	28.87	31.1	34.75	36.01	40.65	47.85	34.96	40.82	43.49	45.05	38.36	43.07	44.4	45.88	45.71
	som	33.9	37.83	39.37	40.67	41.81	46.32	45.39	45.89	45.64	47.73	35.8	47.12	46.8	46.08	50.15
	tir	24.45	24.42	27.86	30.47	31.28	37.6	30.69	32.15	33.11	34.82	29.77	29.05	29.79	30.24	33.69
Prompt 3	amh	25.68	34.1	34.09	35	38	47.03	42.65	41.07	40.61	42.73	39.82	41.93	40.81	42.54	45.17
	orm	27.42	33.17	35.77	37.66	40.61	49.31	39.84	41.38	44.64	44.19	41.82	41.81	41.2	44.75	46.43
	som	31.14	39.9	39.34	42.55	43.02	45.12	46.25	46.1	46.51	47.22	39.5	43.54	44.52	46.22	47.67
	tir	17.6	24.69	27.2	29.02	30.44	35.02	30.85	31.06	33.2	34.02	25.93	27.7	28.88	32.64	35.4

		LLaMA-3-8B-Instruct					LLaMA-3.1-8B-Instruct					CohereForAI__aya-101				
Prompt 1	amh	45.77	47.4	45.87	45.44	46.06	33.85	46.09	47.37	46.16	48.51	42.27	47.81	49.62	49.86	49.27
	orm	47.65	47.96	46.3	46.83	47.38	44.14	48.49	48.41	50.4	49.17	39.86	44.72	45.55	46.66	47.6
	som	42.81	48	48.65	48.39	48.91	36.12	47.04	47.9	49.73	50.66	37.36	46.34	44.48	46.42	47.36
	tir	38	34.13	32.17	34.73	36.34	27.84	33.58	37.05	39.74	40.63	27.01	39.338	38.47	38.08	39.27
Prompt 2	amh	48.12	45.55	44.69	45.31	45.13	38.73	44.51	46.01	45.96	48.25	47.38	48.62	48.72	49.51	47.85
	orm	46.94	47.27	45.64	46.36	47.99	38.5	46.85	46.84	48.62	48.56	43.76	45.71	45.5	47.21	48.03
	som	46.82	46.92	47.09	46.96	48.12	48.03	45.76	47.74	49.69	49.51	41.17	46.83	45.07	46.77	46.77
	tir	40.77	28.77	30.75	34.48	33.25	28.11	30.47	35.8	36.47	38.28	37.5	39.31	38.68	37.08	38.08
Prompt 3	amh	50.89	46.61	44.85	45.04	46.18	13.18	46.07	46.79	46.17	48.08	45.2	46.79	46.54	44.39	44.71
	orm	51.05	45.72	44.78	45.46	46.52	20.65	47.76	48.42	47.3	48.85	46.56	47.93	49	48.19	47
	som	49.75	45.65	47.18	46.94	47.93	20.7	46.89	48.23	48.9	48.48	46.02	46.89	46.16	47.34	47.33
	tir	44.02	33.01	31.51	34.1	37.51	9.02	34.28	37.26	38.66	40.05	30.37	33	32.12	31.44	29.99

Table 9: Prompt sensitivity experiment results across k-shots, LLMs, and languages