EthiopicEmotion: Multi-label Emotion Dataset with Large Language Models Evaluation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) show promising learning and reasoning abilities. Compared to other NLP tasks, multilingual and multi-label emotion evaluation is under-explored in LLMs. In this paper, we present **EthiopicEmotion**, a multi-label emotion with intensity dataset for four Ethiopian languages, namely Amharic (amh), Afan Oromo (oro), Somali (som), and Tigrinya (tir). We perform extensive experiments with additional English multi-label emotion data from SemEval 2018:affect in tweets work. Our evaluation includes zero-shot and in-context learning (ICL) from large language models. The result shows that accurate multi-label emotion classification is still insufficient, even for high-resource language (English), and there is a large gap between the performance of English and low-resource languages.

1 Introduction

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 the author's mental state (A.V. et al., 2024). Although there have been several efforts in constructing emotion benchmark datasets, emotion detection for Ethiopian languages 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). Due to the cultural and language differences inherent in interpreting emotions (Kusal et al., 2023), this work aims to create and evaluate a multi-label emotion with intensity dataset for Ethiopian languages: Amharic (amh), Afan Oromo (oro), Somali (som), and Tigrinya (tir). We follow the multi-label annotation approach because it allows an instance to have any combination (none, one, some, or all) of labels from a given set of emotion labels (Ameer et al., 2020).

2 EthiopicEmotion dataset creation

Data sources: To make the dataset as possible as balanced, we translate the English NRC EmoLex (Mohammad and Turney, 2013) lexicon into Ethiopian languages, and we collect additional emotion lexicons using word similarities from available Word2vec and fastText word embedding models (Yimam et al., 2021; Belay et al., 2021). As rich sources of textual emotions are obtained from social networks such as YouTube, Twitter (X), Facebook, and news headlines, we collected the data from these sources. Table 1 shows the statistics taken from each data source for each language.

Data sources	**amh**	**oro**	**som**	**tir**
Tweeter (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: Twitter (X) posts (tweets), Facebook post comments, YouTube video comments, and news headlines. *Instances that have no label agreement between annotators will be excluded after annotation.*

Data annotation: The annotation process is accomplished by hiring native speakers from each language. We customize the Potato annotation tool (Pei et al., 2022) for our multi-label emotion annotation platform. Each instance is annotated by three native speakers with its corresponding intensity value (0 - for neutral, low, medium, or high). We use Ekman's six basic emotions (anger, disgust, fear, joy, sadness, surprise) plus neutral class (Ekman, 1992). The final label is determined based on agreement by at least two annotators. Final annotated dataset statistics are Amharic 5,891, Afan Oromo 5,690, Somali 5,631, and Tigrinya 6,109, a total of 23,321 instances were annotated.

Model name	Amharic	Afan Oromo	Somali	Tigrinya	English	average
Gemma-2b-i	5.76	2.06	4.91	4.40	32.31	9.88
Gemma-1.1-7b-it	**8.51**	**17.06**	8.99	**6.19**	**49.72**	**18.09**
LLaMA-2-7b-chat-hf	8.00	3.48	**10.48**	3.43	11.56	7.39
LLaMA-3-8B-Instruct	7.86	9.10	4.95	3.77	41.32	13.40

Table 2: Weighted average F1-score results from zero-shot experiments

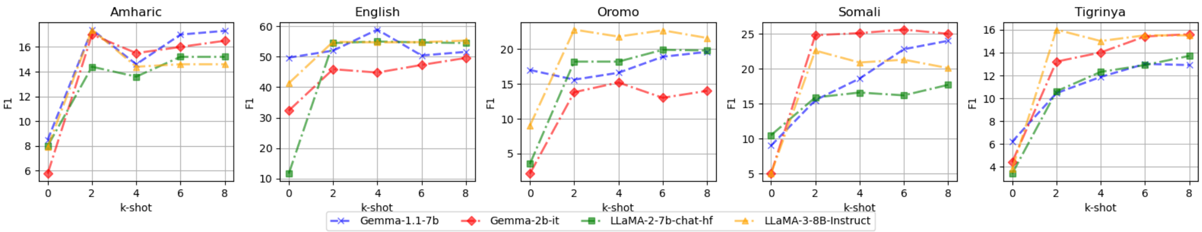

Figure 1: Evaluation result across languages and LLMs with different numbers of shots

3 Inter-Annotator Agreement (IAA)

A measure termed Cohen's kappa (Cohen, 1960), Fleiss' kappa (Fleiss, 1971), Krippendorff's alpha (Krippendorff, 2011), and bootstrapping multiple label agreement (Marchal et al., 2022) are proposed for annotators agreement. However, they don't support multi-label raters and multi-label agreements simultaneously. We adopted a multi-label agreement (MLA) method proposed by Li et al. (2023) to obtain the observed multi-label agreement among multiple raters. The agreement results are 0.50 for Amharic, 0.64 for Afan Oromo, 0.51 for Somali, and 0.53 for Tigrirnya. The IAA result among the seven emotion classes shows moderate agreement.

4 Experiments

To make our evaluation universal, we included the English multi-label emotion dataset from SemEval 2018 Task 1: affect in tweets Mohammad et al. (2018) dataset. For a baseline experiment, we evaluated instruction tuned version of **LLaMA-2-7b-chat-hf** (Touvron et al., 2023), **LLaMA-3-8B-Instruct** (Meta, 2024), **Gemma-2b-i**, and **Gemma-1.1-7b-it** (Team et al., 2024). We selected these generative models because of their popularity in the open-source community, and they are used as a baseline for building similar applications.

5 Results

Zero shot results: The results are presented in Table 2. As we can see, **Gemma-1.1-7b-it** shows consistent top performance in all languages except Somali. For Somali language, **LLaMA-2-7b-chat-hf** performs better result. In a zero-shot experiment, **Gemma-1.1-7b-it** achieves better results for the EthiopicEmotion dataset.

In context learning (ICL): Following the work (Zhang et al., 2024), we leveraged in-context learning to teach the models about the task by showing examples without parameter updates. We worked with 2, 4, 6, and 8 demonstrations in our experiment and compared them with zero-shot experiments as shown in Fig 1. All show improvement in their scores by showing two examples (2 shots) compared to zero-shot tests. However, **Gemma-1.1-7b-it** does not show this improvement in Afan Oromo language, which already had good scores in the zero-shot experiment. When increasing the number of demonstrations from two to eight examples, the improvement is inconsistent and cannot be guaranteed. This is particularly evident in English.

6 Conclusion and Future Work

In this work, we present **EthiopicEmotion** dataset to train and evaluate emotional understanding of models. The dataset provides diversity concerning the source and languages. We have evaluated and reported strong baseline results. We believe that this dataset and experiment results can be employed as a baseline in the future for better multi-label emotion and intensity prediction. As a future work, this work can be extended to experimenting with intensities and evaluating closed-source LLMs. The resources, such as lexicons and datasets, will be publicly available for further investigation.

References

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.

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.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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

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

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

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.

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

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.

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.

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 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. Association for Computational Linguistics.

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.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, and et. al. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

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

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

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 (Online). International Committee on Computational Linguistics.

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. 2024. The impact of demonstrations on multilingual in-context learning: A multidimensional analysis. *arXiv preprint arXiv:2402.12976*.