# Tewodros at SemEval-2025 Task 11: Multilingual Emotion Intensity Detection using Small Language Models

**Mikiyas Mebiratu[2,*], Wendmnew Sitot Abebaw[2,*], Nida Hafeez[1], Fatima Uroosa[1]**
**Tewodros Achamaleh[1], Grigori Sidorov[1], Alexander Gelbukh[1],**
[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico
[2]Wolkite University, Department of Information Technology, Wolkite, Ethiopia
[2]Woldia University, Woldia, Ethiopia
[*]Equal contribution. Corr. email: sidorov@cic.ipn.mx

## Abstract

Emotions play a fundamental role in the decision-making process, shaping human actions across diverse disciplines. The extensive usage of emotion intensity detection approaches has generated substantial research interest during the last few years. Efficient multi-label emotion intensity detection remains unsatisfactory even for high-resource languages, with a substantial performance gap among well-resourced and under-resourced languages. Team **Tewodros** participated in SemEval-2025 Task 11, Track B, focusing on detecting text-based emotion intensity. Our work involved multi-label emotion intensity detection across three languages: Amharic, English, and Spanish, using the (afro-xlmr-large-76L), (DeBERTa-v3-base), and (BERT-base-Spanish-wwm-uncased) models. The models achieved an average F1 score of 0.6503 for Amharic, 0.5943 for English, and an accuracy score of 0.6228 for Spanish. These results demonstrate the effectiveness of our models in capturing emotion intensity across multiple languages.

## 1 Introduction

The modern digital era allows users to freely express their feelings, attitudes, and opinions through websites, microblogs, and social media platforms (Wiebe et al., 2005; Mohammad and Kiritchenko, 2018; Mohammad et al., 2018a; Acheampong et al., 2020; Andalibi and Buss, 2020; Rodríguez-Ibánez et al., 2023). This has increased interest in extracting user sentiments and emotions towards events for different purposes, including social media monitoring, product analysis, political promotions, customer feedback analysis, and marketing research (Nandwani and Verma, 2021; Naidoo et al., 2022; Shehu, 2023; Kusal et al., 2023a; Achamaleh et al., 2025). Language is not just a medium of communication but also a way to express emotions, sentiments, and their intensity (Richards, 2022).

The task of emotion classification within NLP stands as a complex challenge that involves assigning emotional labels to texts to reveal the precise mental state of writers/users (Alswaidan and Menai, 2020; Tao and Fang, 2020; Mohammad, 2022; Saffar et al., 2023). The challenge of emotion detection exceeds sentiment analysis (Birjali et al., 2021) because emotions span a wide spectrum and single texts can contain multiple feelings, while cultural and linguistic differences impact interpretation and transferability (Yu, 2022; Kusal et al., 2023b; Wang et al., 2024b).

Multi-label Emotion Classification (MLEC) (Ameer et al., 2020; Deng and Ren, 2020; Liu et al., 2023) involves analyzing complete emotional expressions within written content, thereby demonstrating its value as a complex yet fundamental Natural Language Processing (NLP) task because one text may convey various simultaneous emotions. Multi-label classification differs from single-label by enabling instances to possess different mixture levels of emotions from the complete emotion set (Belay et al., 2024). Different machine learning (ML) algorithms (Azari et al., 2020; Alslaity and Orji, 2024) such as Naive Bayes (NB), k-nearest neighbors, and Support Vector Machines (SVM) have been applied to resolve emotion classification problems, often incorporating linguistic and contextual features for better performance.

The detection method of emotions in coarse-grained systems only identifies emotions and ignores their intensity level. Traditional emotion classification approaches can determine whether a sentence expresses happiness or sadness but do not quantify how intense the emotion is (Setiawan and Chowanda, 2023). Fine-grained emotion intensity detection aims to capture these variations, which is crucial for distinguishing sentences with the same emotion but different intensities. Detecting emotion intensity requires identifying intensity words and other linguistic factors that influence the



*Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1485–1494

degree of emotion expressed in the text (Mashal and Asnani, 2017; Chutia and Baruah, 2024).

Despite growing research in this domain, most studies focus on data-rich languages due to the unavailability of datasets for data-scarce languages (Mageuresse et al., 2020; Abiola et al., 2025b,a). This gap has led to limited advancements in emotion intensity detection for languages spoken in linguistically diverse regions such as Africa and Asia, which together account for over 4,000 languages (Irwin, 2020; Welmers, 2024).

The workshop organizers launched SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection (Track B) as a response to address the current research gap. The task serves as a research platform that allows scientists and researchers to build and evaluate multi-label emotion intensity detection methods while targeting the gap between languages with varying linguistic resources.

## 2 Literature Review

The process of emotion classification (EC) analyzes verbal and nonverbal indicators, including text and facial expressions, together with body language and speech, to determine a subject's emotional state (Dadebayev et al., 2022; A.V. et al., 2024). The main goal of EC consists of detecting emotions through categorization across text expressions, including anger, disgust, and fear, together with happiness, sadness, and surprise emotions. Psychologists argue about essential emotions, but different psychological frameworks propose between six and twenty emotions as core (Plutchik, 1980; Frijda, 1988; Parrot, 2001; Russell, 2003).

Several dimensional models of emotion have been developed, yet only a limited number persist as dominant frameworks. The Circumplex model (Russell, 2005) features eight emotional groups established through 28 emotion words, and the Positive-and Negative-Activation (PANA) model demonstrates an emotion ranging from high positive to low positive activation (Watson and Tellegen, 1985). Most researchers in emotional-based research continue to use Ekman's model (Ekman, 1992) to divide emotions into six core categories that include joy, surprise, happiness, anger, sadness, disgust, and also fear (Hoemann et al., 2020).

The research community has created text mining solutions to analyze emotions on social media (Goldenberg and Willer, 2023), especially Twitter, through Naïve Bayes machine learning strategies (Wikarsa and Thahir, 2015; Mohammad and Bravo-Marquez, 2017). The tagging of emotions in online news through multi-source systems is addressed by researchers who introduce a two-layer logistic regression model in their approach (Yu et al., 2015; Bostan and Klinger, 2019). Research in emotion classification underwent several developments by integrating word and character n-grams (Mohammad, 2012) with sentiment and emotion lexicons (Mohammad et al., 2015) as well as neural network models (Felbo et al., 2017; Köper et al., 2017).

There has been an increased emphasis in NLP research (Graterol et al., 2021) that incorporates Large Language Models (LLMs) for emotion identification, mainly within data-rich and data-scarce languages (Belay et al., 2024; Muhammad et al., 2025c; Tonja et al., 2024). EmoBench represents a new assessment method that tests LLMs for detecting emotional origins across English and Chinese languages, as described by (Sabour et al., 2024). (Liu et al., 2024) proposed EmoLLMs, fine-tuning open-source LLMs for affective analysis and emotion prediction. The researchers (Cageggi et al., 2023) applied MT5 model fine-tuning before conducting evaluations of both FLAN and ChatGPT through few-shot prompting for multi-label emotion classification.

Mostly used emotion classification datasets exist, including: (Wang et al., 2024a) SemEval-2024 Task 3, (Muhammad et al., 2025c) SemEval Task 11, (Muhammad et al., 2025b,a) BRIGHTER, (Bianchi et al., 2022) Multilingual Emotion Prediction (XLM-EMO), (Ameer et al., 2023) WASSA 2023 Shared Task 2, (Ciobotaru et al., 2022) Romanian Emotion Dataset (REDv2), (Demszky et al., 2020) GoEmotions, and Balanced Multi-Label Emotional Tweets (BMET) (Huang et al., 2021).

The detection of emotion intensity in text (Zad et al., 2021) has become a core capability of the multi-label emotion classification (MLEC) process to identify emotional levels (Ameer et al., 2020). The SemEval-2018 Task 1 (Mohammad et al., 2018b) together with the Multimodal Multi-label Emotion, Intensity, and Sentiment Dialogue Dataset (MEISD) (Firdaus et al., 2020) and EmoIn-Hindi (Singh et al., 2022) demonstrate examples of emotion classification.

The development of emotion analysis through deep learning techniques signifies an increasing preference for sophisticated emotion detection methods. Distinguishing sentences from the same

emotion category requires examining their emotional strength because intensity plays a vital role in these cases (Htait et al., 2016; Refaee and Rieser, 2016; Lenc et al., 2016). The intensity of individual words stands as an approach to measure sentence-level intensity since words that share related meanings push emotional strength either upward or downward (Alejo et al., 2020).

The detection of emotional intensity in text documents remains a subject that researchers have studied comparatively less than sentiment intensity (Alm et al., 2005; Aman and Szpakowicz, 2007; Bollen et al., 2009; Neviarouskaya et al., 2009; Brooks et al., 2013). This gap in research has been addressed through several notable studies. The semi-supervised approach defines an adjective intensity scale through contextual analysis of high-intensity words (Sharma et al., 2015). Sciences have studied automated approaches for emotional intensity tagging in sentences with WordNet Affect alongside word sense disambiguation (de Albornoz et al., 2010). These research techniques aim to establish quantitative methods that identify emotional intensity levels across a given textual content.

The first major attempt to introduce emotion intensity annotation occurred when (Strapparava and Mihalcea, 2007) participated in SemEval-2007 shared task competitions. The study employed a 0 to 100 continuous scale through which annotators rated emotions present in newspaper headlines. The development of rating emotion intensity at a granular level still encounters specific collection difficulties. The process faces major problems because different annotators tend to rate the same piece of text with substantially varying scores (one person assigned 79 while another only gave 62).

Researchers have made important progress in emotion classification, but their work mostly concentrates on high-resource languages (Strapparava and Mihalcea, 2007; Seyeditabari et al., 2018; Chatterjee et al., 2019; Kumar et al., 2022), whereas the investigation of emotion detection within Ethiopian languages along with other low-resource languages remains scarce (Muhammad et al., 2025b). Benchmark datasets primarily emerge for English together with popular languages, which impede proper generalization of research outcomes across diverse linguistic settings (Yimam et al., 2020; Tela et al., 2020; Muhammad et al., 2023). The current emotion datasets derive from single-source text corpora, which affect their representativeness, while state-of-the-art LLMs for both multi-label and multilingual emotion classification remain underexplored domains (Yimam et al., 2021). Our team extends this research on Track B of SemEval 2025 Task 11 to address the existing gaps in emotion intensity detection. The assessment task requires emotion annotation with perceived levels of intensity, which range from 0 for no emotion to 3 for high intensity according to sadness, fear, and so on. To detect emotion intensity in high- and low-resource languages, this work investigates methods and datasets and demonstrates results. Extending from the extant literature, it is our intention to contribute by culturally informed approaches to improve the efficiency of this task.

## 3 Methodology

The research methodology included data preprocessing, feature extraction, and text classification procedures on textual data collected from diverse sources in English, Spanish, and Amharic languages. The data was divided into three distinct sets: training, development, and testing. Each sample was labeled with one of six emotional states: surprise, anger, joy, fear, disgust, joy, and sadness. We loaded the data through Pandas to examine its structure while confirming essential attributes exist. We determined emotion frequencies through value_count functions.

The feature engineering process involved CountVectorizer's bigram tokenization technique alongside a selection of the top 90 bigrams per language for improving input representation. The model design included provisions that allowed it to detect important linguistic patterns regardless of the text's language.

The model architecture incorporated Afro-XLMR-Large-76L for Amharic, DeBERTa-v3-Base for English, and BERT-Base-Spanish-WWM-Uncased for Spanish. Different tokenization methods applied to textual data through model-specific tokenizers resulted in Hugging Face datasets for training purposes. The training speed was accelerated by using fp16=True during mixed precision operations. The training process employed 16 samples per batch and set the learning rate at 2e-5. A Trainer class and CrossEntropyLoss module enabled the computation of class weights for balancing class distribution in the training process. The training process lasted for 20 epochs through early stopping to avoid overfitting. Multiple performance metrics, including accuracy, macro F1-score, and

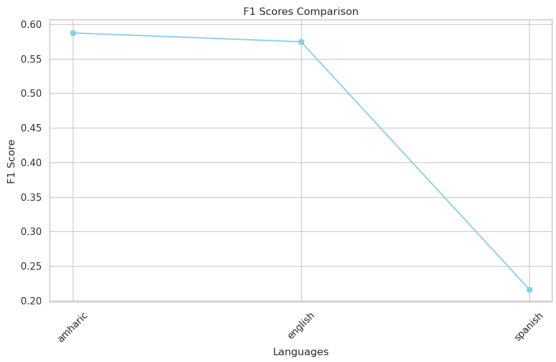


Figure 1: F1 Comparison Scores of English, Spanish and Amharic Languages from the Dev Set

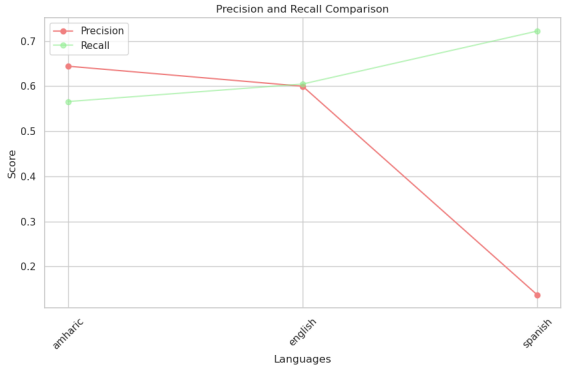


Figure 2: Precision and Recall Scores Comparison of Amharic, English and Spanish Languages from the Dev Set

per-label F1-score, served to assess model performance during evaluation.

## 4 Dataset Analysis

The SemEval 2025 Task 11 dataset, Track B (Emotion Intensity Detection), originates from the paper (Belay et al., 2024). The dataset provides labeled intensity scores for joy, sadness, fear, anger, surprise, and disgust. Our analysis focuses on examining three languages: English, Spanish, and Amharic, which represent a mix of high-resource and low-resource linguistic contexts. The dataset is compiled from diverse sources, including news portals X (formerly Twitter), YouTube, and Facebook, capturing a mix of formal and informal emotional expressions. Social media data poses distinct challenges, such as the use of abbreviations, informal phrasing, and extreme variations in emotional intensity, while news articles typically present structured and neutral emotional tones. Table 1 demonstrats the emotion distribution across datasets (Train and Dev) for Spanish, English, and Amharic, illustrating the varying intensity of all 6 emotional stats. Through this analysis, we seek to address the gap in emotion intensity detection for both data-rich and data-scarce languages, ensuring that models effectively process diverse grammatical structures and cultural nuances in emotional expression.

## 5 Experimental Setup

Experimental Setup Our experiment utilized a training-validation split on the dataset to ensure a balanced distribution of emotion intensity labels for the six perceived emotions: joy, sadness, fear, anger, surprise, and disgust. Language-specific preprocessing techniques were applied to address platform-specific variations and the models were trained in a high-performance computing environment for efficient training. For model training, we fine-tuned Afro-XLMR-Large-76L, DeBERTa-v3-Base, and BERT-Base-Spanish-WWM-Uncased, leveraging their capabilities in multilingual and language-specific emotion detection. Each model was initialized with pre-trained weights and fine-tuned on the dataset to capture nuanced emotional patterns across the six emotion categories. To evaluate model performance, we developed a comprehensive evaluation pipeline, using accuracy, F1-score, and Pearson correlation as key metrics to assess the effectiveness of emotion intensity detection. Generalization was tested on unseen test data from different platforms and contexts. Additionally, we also integrated language-agnostic embeddings to enhance robustness across multiple languages. Model hyperparameters were optimized through experimental tuning to balance precision and computational efficiency in detecting joy, sadness, fear, anger, surprise, and disgust across varied textual contexts.

## 6 Results

Our evaluation focuses on measuring the model's effectiveness in detecting emotion intensity across the three languages: English, Spanish, and Amharic. We used Pearson correlation (r) as a key metric to evaluate the alignment between predicted emotion intensity scores and gold-standard annotations. Table 2 summarizes the model's performance across the six emotions: joy, surprise, fear, anger, disgust, and sadness based on the test set that the workshop organizer's provided. Overall, the model achieved its highest performance in Amharic, followed by Spanish and English, suggesting its abil-

| Dataset | Anger | Disgust | Fear | Joy | Sadness | Surprise | Total |
|---|---|---|---|---|---|---|---|
| Spanish (Train) | 939 | 1343 | 635 | 1315 | 635 | 840 | 5707 |
| Spanish (Dev) | 68 | 136 | 63 | 115 | 59 | 83 | 524 |
| English (Train) | 497 | 0 | 2573 | 963 | 1376 | 1126 | 6535 |
| English (Dev) | 27 | 0 | 96 | 43 | 54 | 43 | 263 |
| Amharic (Train) | 1429 | 1878 | 145 | 883 | 1211 | 165 | 5711 |
| Amharic (Dev) | 234 | 310 | 22 | 147 | 203 | 31 | 947 |

Table 1: Emotion distribution across datasets (Train and Dev) for Spanish, English, and Amharic.

| Language | Anger | Disgust | Fear | Joy | Sadness | Surprise | Avg. Pearson r |
|---|---|---|---|---|---|---|---|
| Amharic | 0.5406 | 0.6775 | 0.5656 | 0.7997 | 0.7343 | 0.5843 | 0.6503 |
| English | 0.4761 | - | 0.6606 | 0.7276 | 0.6162 | 0.4909 | 0.5943 |
| Spanish | 0.5942 | 0.6180 | 0.6764 | 0.6246 | 0.6114 | 0.6124 | 0.6228 |

Table 2: Emotion Scores across different Languages

| Language | Model | F1 Score |
|---|---|---|
| Amharic | afro-xlmr-large-76L | 0.6503 |
| English | DeBERTa-v3-base | 0.5943 |
| Spanish | spanish-wwm-uncased | 0.6228 |

Table 3: Results obtained from the testset for emotion intensity.

ity to capture emotion intensity variations even in a low-resource language. These findings highlight the model's capability to generalize across different languages, despite variations in linguistic resources and data availability. Figures 1 and 2 highlight the F1, precision, and recall scores for all three languages, while Table 3 presents the F1 score for each language.

## 7 Discussion

The F1 Scores of AfroXLMR-Base reached their highest levels across all languages, especially Amharic and Spanish, which proves its powerful multi-language capabilities. DeBERTa succeeded within English datasets but experienced difficulties working with languages with fewer available resources. The BERT-based Spanish model performed effectively for Spanish tasks but displayed a weak translation ability between languages. The models' performance suffered mostly because of class imbalance when identifying rare emotions, including disgust and surprise. Transformer models demonstrated superior performance than traditional deep learning approaches, as AfroXLMR achieved the best results in precision and recall metrics. The research agenda should encompass emotion-based pre-training and approaches to address imbalanced classes. Table 3 summarizes the model results.

### 7.1 Error Analysis

Most misclassifications happened in Amharic, indicating difficulties with low-resource languages and class imbalance issues. The best F1 score of AfroXLMR could not prevent it from mistaking emotions with similar intensity levels, particularly between sadness and fear. DeBERTa made frequent mistakes in English by labelling strong emotions as moderate since they were written with subtle indicators. The Spanish predictions displayed misinterpretations between joy and surprise categories due to patterns that were similar in the language. The detection accuracy needs better fine-tuning of models alongside additional emotional features and balanced training datasets to achieve superior outcomes in multilingual emotion inscription detection.

## 8 Conclusion

This research work used transformer models that were fine-tuned specifically for Amharic, English, and Spanish to understand emotional intensity across the three languages. The research process included text preprocessing along with feature engineering before training Afro-XLMR-Large-76L, DeBERTa-v3-Base, and BERT-Base-Spanish-WWM-Uncased models. The tested models exhibited superior performance for detecting emotion intensity variations in Amharic with an F1 score of 0.6503, followed by Spanish with an F1 score of 0.6228, and English with an F1 score of 0.5943. While our models perform well across multiple

languages, detection of underrepresented emotions and low-resource languages remains a challenge. Future work should focus on exploring data augmentation techniques and developing adaptation frameworks to achieve better results with multilingual fusion approaches. These advancements will further enhance emotion intensity detection, ensuring more robust and accurate predictions across diverse linguistic and cultural contexts.

## Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT 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.

## References

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Oluwatobi Joseph Abiola, Temitope Olasunkanmi Oladepo, Olumide Ebenezer Ojo, Grigori Sidorov, and Olga Kolesnikova. 2025a. CIC-NLP at GenAI detection task 1: Leveraging DistilBERT for detecting machine-generated text in English. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 271–277, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Olumide Ebenezer Ojo. 2025b. CIC-NLP at GenAI detection task 1: Advancing multilingual machine-generated text detection. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 262–270, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Tewodros Achamaleh, Nida Hafeez, Mikiyas Mebraihtu, Fatima Uroosa, and Grigori Sidorov. 2025. CIC-NLP@DravidianLangTech-2025: Fake News Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech-2025) at NAACL 2025*. Association for Computational Linguistics. To appear.

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

Irean Navas Alejo, Toni Badia, and Jeremy Barnes. 2020. Cross-lingual emotion intensity prediction. *arXiv preprint arXiv:2004.04103*.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Joint Conference on HLT–EMNLP*, Vancouver, Canada.

Alaa Alslaity and Rita Orji. 2024. Machine learning techniques for emotion detection and sentiment analysis: Current state, challenges, and future directions. *Behaviour Information Technology*, 43(1):139–164.

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Text, Speech and Dialogue*, volume 4629 of *Lecture Notes in Computer Science*, pages 196–205. Springer.

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.

Nazanin Andalibi and Justin Buss. 2020. The human in emotion recognition on social media: Attitudes, outcomes, risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–16. ACM.

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.

Bahar Azari, Christiana Westlin, Ajay B. Satpute, J. Benjamin Hutchinson, Philip A. Kragel, Katie Hoemann, Zulqarnain Khan, et al. 2020. Comparing supervised and unsupervised approaches to emotion categorization in the human brain, body, and subjective experience. *Scientific Reports*, 10(1):20284.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philipp Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2024. Evaluating the capabilities of large language models for multi-label emotion understanding. *arXiv preprint*.

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.

Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

Johan Bollen, Huina Mao, and Alberto Pepe. 2009. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, pages 450–453.

Laura Ana Maria Bostan and Roman Klinger. 2019. Exploring fine-tuned embeddings that model intensifiers for emotion analysis. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 25–34, Minneapolis, USA. Association for Computational Linguistics.

Michael Brooks, Katie Kuksenok, Megan K. Torkildson, Daniel Perry, John J. Robinson, Taylor J. Scott, Ona Anicello, Ariana Zukowski, and Harris. 2013. Statistical affect detection in collaborative chat. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 317–328, San Antonio, Texas, USA.

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.

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

Tulika Chutia and Nomi Baruah. 2024. Text-based emotion detection: A review. *International Journal of Digital Technologies*, 3(I).

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.

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.

Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervás. 2010. Improving emotional intensity classification using word sense disambiguation. *Research in Computing Science*, 46:131–142. Special Issue: Natural Language Processing and its Applications.

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.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3):169–200.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. 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.

Nico H. Frijda. 1988. The laws of emotion. *American Psychologist*, 43(5):349.

Amit Goldenberg and Robb Willer. 2023. Amplification of emotion on social media. *Nature Human Behaviour*, 7(6):845–846.

Wilfredo Graterol, Jose Diaz-Amado, Yudith Cardinale, Irvin Dongo, Edmundo Lopes-Silva, and Cleia Santos-Libarino. 2021. Emotion detection for social robots based on nlp transformers and an emotion ontology. *Sensors*, 21(4):1322.

Katie Hoemann, Rachel Wu, Vanessa LoBue, Lisa M. Oakes, Fei Xu, and Lisa Feldman Barrett. 2020. Developing an understanding of emotion categories: Lessons from objects. *Trends in Cognitive Sciences*, 24(1):39–51.

Amal Htait, Sebastien Fournier, and Patrice Bellot. 2016. Lsis at semeval-2016 task 7: Using web search engines for english and arabic unsupervised sentiment intensity prediction. In *Proceedings of SemEval-2016*, pages 469–473, San Diego, California.

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.

Harry Irwin. 2020. *Communicating with Asia: Understanding People and Customs*. Routledge.

Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems*, 240:108112.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2023a. 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.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2023b. 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.

Maximilian Köper, Evgeny Kim, and Roman Klinger. 2017. Ims at emoint-2017: Emotion intensity prediction with affective norms, automatically extended resources and deep learning. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–57, Copenhagen, Denmark.

Ladislav Lenc, Pavel Král, and Václav Rajtmajer. 2016. Uwb at semeval-2016 task 7: Novel method for automatic sentiment intensity determination. In *Proceedings of SemEval-2016*, pages 481–485, San Diego, California.

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. *arXiv preprint*.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Sonia Xylina Mashal and Kavita Asnani. 2017. Emotion intensity detection for social media data. In *Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication*.

Saif Mohammad. 2012. emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018a. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018b. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, LA, USA.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Preprint, arXiv:2109.08256*.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. *arXiv preprint*.

Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.

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.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao,

Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhangand Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, et al. 2025b. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv preprint arXiv:2502.11926*.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025c. 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.

Shaldon Wade Naidoo, Nalindren Naicker, Sulaiman Saleem Patel, and Prinavin Govender. 2022. Computer vision: the effectiveness of deep learning for emotion detection in marketing campaigns. *International Journal of Advanced Computer Science and Applications*, 13(5).

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2009. Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM-09)*, pages 278–281, San Jose, California.

W. Parrot. 2001. *Emotions in Social Psychology*. Psychology Press.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, Research, and Experience*, 1(3):3–33.

Eshrag Refaee and Verena Rieser. 2016. ilab-edinburgh at semeval-2016 task 7: A hybrid approach for determining sentiment intensity of arabic twitter phrases. In *Proceedings of SemEval-2016*, pages 474–480, San Diego, California.

Jack C. Richards. 2022. Exploring emotions in language teaching. *RELC Journal*, 53(1):225–239.

Margarita Rodríguez-Ibánez, Antonio Casánez-Ventura, Félix Castejón-Mateos, and Pedro-Manuel Cuenca-Jiménez. 2023. A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, 223:119862.

James A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):1–145.

James A. Russell. 2005. Emotion in human consciousness is built on core affect. *Journal of Consciousness Studies*, 12(8-10):26–42.

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.

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

Rindy Claudia Setiawan and Andry Chowanda. 2023. Emotion intensity value prediction with machine learning approach on twitter. *CommIT (Communication and Information Technology) Journal*, 17(2):235–243.

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. 2018. Emotion detection in text: A review. *arXiv preprint*.

Raksha Sharma, Mohit Gupta, Astha Agarwal, and Pushpak Bhattacharyya. 2015. Adjective intensity and sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2520–2526.

Harisu Abdullahi Shehu. 2023. *Emotion Detection and its Effect on Decision-making*. Phd dissertation, Te Herenga Waka-Victoria University of Wellington.

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.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of SemEval-2007*, pages 70–74, Prague, Czech Republic.

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

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, et al. 2024. Ethiollm: Multilingual large language models for ethiopian languages with task evaluation. *arXiv preprint arXiv:2403.13737*.

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*.

David Watson and Auke Tellegen. 1985. Towards a consensual structure of mood. *Psychological Bulletin*, 98(2):219–235.

Wm E. Welmers. 2024. *African Language Structures*. University of California Press.

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

Liza Wikarsa and Sherly Novianti Thahir. 2015. A text mining application of emotion classification of twitter users using naïve bayes method. In *IEEE Conference*.

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).

Li Yu, Zhifan Yang, Peng Nie, Xue Zhao, and Ying Zhang. 2015. Multi-source emotion tagging for online news. In *12th Web Information System and Application Conference*.

Qiangfu Yu. 2022. A review of foreign language learners' emotions. *Frontiers in Psychology*, 12:827104.

Samira Zad, Maryam Heidari, H. James Jr, and Ozlem Uzuner. 2021. Emotion detection of textual data: An interdisciplinary survey. In *2021 IEEE World AI IoT Congress (AIIoT)*, pages 0255–0261. IEEE.