NUST Titans at SemEval-2025 Task 11: AfroEmo: Multilingual Emotion Detection with Adaptive Afro-XLM-R

Mehwish Fatima[1], Maham Khan[1], Hajra Binte Naeem[1], Faiza Khan[1], Laiba Rana[1], Seemab Latif[1], and Raja Khurram Shahzad[2]*

[1]School of Electrical Engineering and Computer Science,
National University of Sciences and Technology, Islamabad, Pakistan
[2]Department of Communication, Quality Management and Information Systems,
Mid Sweden University, Sweden
{mehwish.fatima, seemab.latif}@seecs.edu.pk, raja-khurram.shahzad@miun.se

Abstract

Emotion detection in text is particularly challenging for low-resource languages due to linguistic diversity and cultural nuances. To promote inclusivity, SemEval introduces a multilingual, multi-label emotion dataset spanning several low-resource languages. We analyze this dataset to examine cross-lingual emotion distribution and address the performance limitations of existing models, which often overfit to high-resource language patterns. We propose AfroEmo, a multilingual emotion classification model built on Afro-XLM-R. Our approach involves adaptive pre-training on domain-specific corpora, followed by fine-tuning on the shared task dataset. We evaluate our model using Macro-F1, Micro-F1, and other official metrics. AfroEmo achieves a Macro-F1 of 0.71 on Amharic and shows strong generalization to Hausa and Yoruba. We further conduct an ablation study and error analysis to assess the contributions of each model component.

1 Introduction

Emotion detection refers to the process of identifying and interpreting human emotions by analyzing indicators such as body language, tone of voice, facial expressions, and physiological signals. It has broad applications across multiple domains, such as in mental health for detecting depression and emotional distress (Calvo et al., 2015; Baziotis et al., 2018; Almutairi et al., 2024), enhancing customer service engagement (Cambria et al., 2017; Poria et al., 2019), improving cross-cultural communication (Colombo et al., 2020) and in conversational agents to create more emotionally intelligent interactions (Kusal et al., 2024). Traditionally, emotion detection is performed in a monolingual setting. However, researchers have attempted to shift the paradigm toward multilingual emotion detection. Shifting from monolingual to multilingual emotion detection poses challenges, including data scarcity, linguistic diversity, reduced model interpretability across languages (Wang et al., 2024; De Bruyne, 2023; Zhang et al., 2024) and the complexities of code-switching (Wang et al., 2024; De Bruyne, 2023; Zhang et al., 2024). For example, detecting sentiment in an Urdu-English code-mixed sentence like '*I am feeling so udaas today*" ("*udaas*" meaning "*sad*" in Urdu) requires nuanced interpretation that many models lack. Low-resource languages exacerbate these issues (Muhammad et al., 2023), as existing pre-trained models often focus on resource-rich languages (Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020), thus, failing to capture emotional subtleties in underrepresented languages (Tatariya et al., 2023). To address key challenges, SemEval-2025 Task 11 (Muhammad et al., 2025b) introduces a large-scale, low-resource multilingual emotion dataset. This dataset includes 32 languages from seven language families, featuring many underrepresented languages from Africa, Asia, and Latin America. It contains over 100,000 instances, manually annotated by native speakers across six emotion classes: Joy, Sadness, Fear, Anger, Surprise, and Disgust, with emotion intensity on a 4-point Likert scale (0 to 3). Consequently, in this work, we present the following contributions:

- An exploratory data analysis is performed to examine the distribution of emotional indices across multiple low-resource languages.
- We develop a robust multilingual emotion detection model for low-resource languages based on Afro-XLM-R. Our model consists of a two-stage process: adaptive pre-training to improve low-resource language understanding, followed by fine-tuning with the multilingual emotion dataset. The source code is publicly available on GitHub.
- Our empirical evaluation shows strong performance on the test set, ranking among the top

*Corresponding author: raja-khurram.shahzad@miun.se

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2225–2232

systems on the benchmark.

- We also conduct error analysis and an ablation study assessing model's performance and two-stage process.

2 Related Work

2.1 Monolingual

Monolingual emotion detection focuses on identifying emotions within a single language. Supervised and lexicon-based methods perform well on labeled datasets, especially when paired with techniques like TF-IDF and word embeddings (Malagi et al., 2023). Deep learning plays a central role, with convolutional neural networks (CNNs) capturing emotion-bearing phrases and Bidirectional LSTMs (BiLSTMs) enhancing contextual understanding (Trimukhe et al., 2024; V and K J, 2024). Integrating BERT with BiLSTM or parallel CNN blocks further improves performance by leveraging contextual and bidirectional representations. Psycholinguistic features also contribute to higher accuracy in monolingual settings (Juyal and Kundalya, 2023).

2.2 Multilingual

Recent work increasingly targets multilingual emotion detection in low-resource languages. Muhammad et al. (2023) introduce AfriSenti-SemEval, while Raihan et al. (2024) present EmoMix-3L, highlighting the value of pre-trained models like MuRIL. Ameer et al. (2023) propose a multi-attention RoBERTa model for multi-label emotion classification. Despite progress, transformer models such as Afro-XLM-R and XLM-RoBERTa still struggle with emotional subtleties. To address this, researchers improve transfer learning using models like BERT and GoEmotions, and explore cultural context to enhance accuracy (Barnes et al., 2022). Other efforts focus on automatic feature selection (Haider et al., 2021) and evaluating large language models for multilingual, multi-label emotion tasks (Belay et al., 2025).

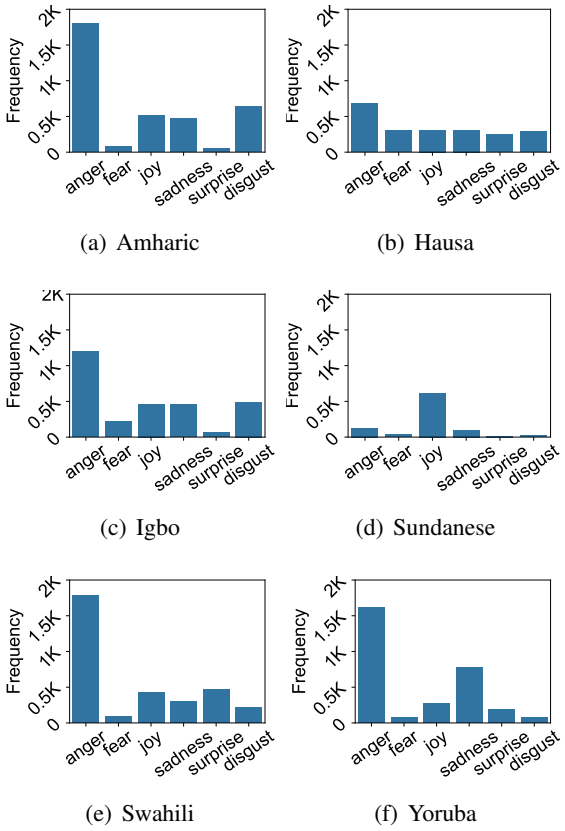
In summary, while traditional methods offer interpretability, deep learning—especially transformer-based models dominates monolingual emotion detection. In multilingual contexts, transfer learning and curated datasets continue to advance the field, though challenges with nuanced emotions and domain variability remain.

Language	Train	Dev	Test	Total
Hausa (hau)	2,145	356	1,080	3,581
Igbo (ibo)	2,880	479	1,444	4,803
Sundanese (sun)	924	199	926	2,049
Swahili (swa)	3,307	551	1,656	5,514
Yoruba (yor)	2,992	497	1,500	4,989

Table 1: Overview of the multilingual emotion dataset.

3 Dataset Analysis

We perform an exploratory data analysis to gain insights of distribution of emotional indices across languages. Table 1 summarizes the dataset properties Hausa, Igbo, Swahili, Sundanese, and Yoruba (Muhammad et al., 2025a). The number of annotators per sample varies slightly, with social media and news articles serving as the primary data sources.



(a) Amharic (b) Hausa

(c) Igbo (d) Sundanese

(e) Swahili (f) Yoruba

Figure 1: Emotion distribution in all six languages

Figure 1 illustrates the distribution of six emotion categories—*anger, joy, sadness, fear, disgust, and surprise*—across six low-resource languages. A strong class imbalance is evident, with *anger* and *joy* dominating most languages, particularly Amharic, Igbo, and Yoruba. In contrast, Hausa exhibits a relatively more balanced distribution. Sundanese stands out with *joy* as the most prevalent

emotion. The limited representation of *surprise*, *disgust* and *fear* across most languages highlights the challenge of developing robust multilingual emotion classifiers, especially under low-resource and imbalanced conditions.

4 Proposed Model

We propose AfroEmo, a multilingual emotion detection model built upon the Afro-XLM-R Large architecture (Alabi et al., 2022), which comprises 24 transformer layers, each with 16 self-attention heads and a 4096-dimensional feedforward network. As shown in Figure 2, our training process consists of two stages: adaptive pre-training on domain-specific corpora and fine-tuning on a labeled multilingual emotion dataset. The implementation and data are publicly available on GitHub.[1]. Hereafter, we call our proposed model as AfroEmo.

4.1 Domain Corpus

We use a diverse set of corpora sourced from Hugging Face, focusing on low-resource African languages across a variety of domains including folklore, conversational text, and web-scraped content. We consider the following datasets: KerVerse (2023) for Amharic, Gurgurov (2023c) for Swahili, Gurgurov (2023b) for Sundanese, Gurgurov (2023a) for Igbo, and Babs (2023) for Hausa. Due to hardware limitations, we randomly sample approximately 30% of each corpus per epoch for adaptive pre-training. This sampling strategy enables efficient training while preserving linguistic diversity across domains. By resampling every epoch, we ensure sufficient exposure to the broader language distribution without overwhelming compute resources.

4.2 Adaptive Pre-training

In the first stage, we pretrain Afro-XLM-R using the masked language modeling (MLM) objective to adapt it to the domain-specific corpora. We apply a 15% token masking strategy, enabling the model to better capture contextual semantics in low-resource language settings. Inputs are tokenized and embedded before passing through the model's 24 transformer layers. The architecture includes a pooling layer, a feedforward layer with ReLU activation, a dense output layer matching vocabulary size, and a softmax layer to produce token probabilities.
We train the model for three epochs and evaluate it on domain-representative validation samples to ensure effective adaptation. This Pre-training phase aims to bridge the domain gap between the original training data and our target emotion detection task.

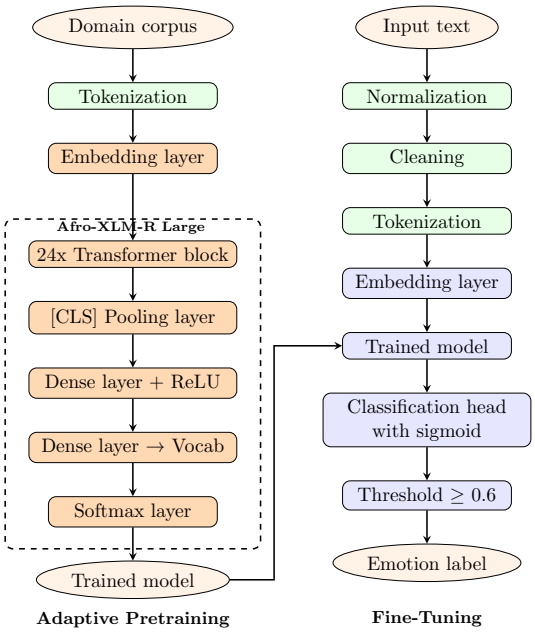


Figure 2: The base model for AfroEmo is Afro-XML-R. The first stage involves adaptive pre-training on domain-specific corpora. The second stage fine-tunes the model for multi-label multilingual emotion classification.

4.3 Preprocessing

Prior to fine-tuning, we preprocess the labeled emotion dataset. We perform the following preprocessing: (1) convert text to lowercase, (2) normalize non-ASCII characters to their ASCII equivalents, (3) remove emojis, special characters, and stopwords, and (4) apply subword tokenization to handle rare or out-of-vocabulary words. This preprocessing pipeline ensures clean, standardized input that emphasizes semantically meaningful content.

4.4 Fine-tuning

We fine-tune the adaptively pretrained model for multilingual, multi-label emotion classification. Emotion labels are converted to binary vectors to support multi-label learning. The classification head consists of two feedforward layers ($1024 \rightarrow 512$ with ReLU, followed by $512 \rightarrow 5$), a sigmoid activation function, and a final thresholding step (≥ 0.6) to produce binary emotion predictions. The model classifies each input into one or more of the five target emotion categories: *joy*, *anger*, *surprise*, *disgust*, and *sadness*.

[1] https://github.com/mhm930/NustTitans

5 Experimental Setup

5.1 Dataset

We adopt the official dataset splits provided by the SemEval-2025 Task 11 organizers. For each language (except Sundanese), the data is partitioned into 60% training, 10% development, and 30% test samples for each language except Sundanese. As Sundanese contains fewer samples (2049), it is split into 45% training, 10% development, and 45% test samples. We use the train and development sets for evaluating and tuning our model, while the final evaluation is conducted on the unseen test set.

5.2 Models

We compare our model with the following baselines: (1) LaBSE (Feng et al., 2022), (2) RemBERT (Chung et al., 2020), (3) XLM-R (Conneau et al., 2020), (4) mBERT (Devlin et al., 2019), and (5) mDeBERTa (He et al., 2020).

5.3 Evaluation

We evaluate the perfomance with given metrics by SemEval shared task: Macro F1-score, Micro F1-score, precision, recall, and accuracy. Macro F1 accounts for per-class performance, while Micro F1 is sensitive to class imbalance—making both essential for evaluating multi-label classification in low-resource settings.

5.4 Training Details

We conduct systematic experiments to identify optimal hyperparameters for both adaptive pre-training and fine-tuning phases. For adaptive pre-training, we use a learning rate of 2×10^{-5}, a training batch size of 8, an evaluation batch size of 16, and a maximum sequence length of 128 tokens for 3 epochs. We use a 15% token masking rate for the masked language modeling (MLM) objective and monitor validation loss to ensure effective domain adaptation.

The configuration settings for fine-tuning are as follows: a learning rate of 5×10^{-6}, a batch size of 8, and a sequence length of 128 tokens over the course of 20 epochs. We use binary cross-entropy loss as our training objective with a sigmoid activation function to enable the prediction of multiple emotion labels for each instance.

5.5 Libraries and Hardware

All experiments are conducted on Kaggle's cloud-based infrastructure, utilizing NVIDIA Tesla P100/T4 GPUs with 16GB RAM. The experiments are implemented in Python by using the following libraries. For preprocessing the input text, we use re, nltk, unidecode, sentencepiece, and nltk stopwords libraries. To transform the output, we use sklearn multi label binarizer. For adaptive pre-training and fine-tuning, we use Hugging Face's Trainer API.

Lang	Mic-F1	Mac-F1	Acc	P	R	Rank
Amh	0.90	**0.71**	0.53	0.70	0.73	**2**
Hau	0.88	**0.69**	0.51	0.68	0.71	**4**
Ibo	0.69	0.18	0.13	0.13	0.14	31
Yor	0.91	0.31	0.58	0.38	0.29	10
Swa	0.85	0.29	0.38	0.29	0.31	15
Sun	0.83	0.35	0.42	0.41	0.33	27
Overall	**0.84**	**0.42**				40

Table 2: Emotion classification results with Mac(ro)-F1, Mic(ro)-F1, Acc(uracy), P(recision), R(ecall) and Rank on the test dataset.

6 Results

We evaluate the performance of AfroEmo against multilingual baselines using standard metrics given by organizers.

6.1 Overall Performance

Table 2 summarizes AfroEmo's performance on the multilingual test set across a diverse set of languages, including both African and non-African. Among African languages, Amharic achieves the strongest results, attaining a Macro-F1 score of 0.71 and ranking second overall across all languages. Hausa follows with balanced metrics and ranks fourth, while Yoruba demonstrates moderate performance, particularly in precision (0.58) and joy detection (0.38), placing 13th. In contrast, Igbo and Sundanese show significantly lower recall and F1 scores, highlighting difficulties in generalization. Swahili falls in the mid-range, ranked 19th, suggesting partial adaptation.

The disparity between Macro-F1 and Micro-F1 scores reveals AfroEmo's sensitivity to class imbalance. While high Micro-F1 scores suggest the model captures dominant emotional expressions well, lower Macro-F1 scores across languages reflect its difficulty in detecting minority classes consistently—an ongoing challenge in multilingual, multi-label emotion classification.

6.2 Comparison with Baselines

Table 3 presents a comparison of the Macro-F1 scores achieved by AfroEmo and five baseline mod-

Model	Hau	Ibo	Yor	Swa	Sun	Avg.
LaBSE	0.38	0.18	0.11	0.21	**0.35**	0.25
RemBERT	0.31	0.74	0.53	0.19	0.19	**0.39**
XLM-R	0.16	0.10	0.66	0.17	0.26	0.27
mBERT	0.15	**0.99**	**0.96**	0.18	0.25	0.5
mDeBERTa	0.32	0.95	0.10	0.15	0.27	0.36
AfroEmo	**0.69**	0.18	0.31	**0.29**	**0.35**	0.36

Table 3: Macro-F1 comparison of AfroEmo with LaBSE, RemBERT, XLM-R, mBERT, mDeBERTa on five languages.

Emotion	Amh	Hau	Ibo	Yor	Swa	Sun
Anger	**0.67**	0.62	0.16	0.39	0.30	0.17
Disgust	0.77	**0.79**	0.23	0.13	0.21	0.19
Fear	0.64	**0.75**	0.04	0.00	0.17	0.00
Joy	**0.77**	0.71	0.40	0.38	0.41	0.83
Sadness	**0.75**	0.72	0.23	0.68	0.32	0.57
Surprise	**0.68**	0.57	0.02	0.30	0.35	0.32

Table 4: Emotion-wise Macro-F1 on the test dataset for all languages.

els: LaBSE, RemBERT, XLM-R, mBERT, and mDeBERTa. The proposed AfroEmo model outperforms the baselines in three languages—Hausa, Swahili, and Sundanese. While the average Macro-F1 score of AfroEmo is comparable to that of mDeBERTa, RemBERT achieves the highest overall performance. It is noteworthy that several of these baseline models are not evaluated on Amharic due to Vocabulary Limitation, as it is an underrepresented language, the best-performing low-resource language for AFroEmo. To improve the Macro-F1 score, class imbalance is addressed using techniques such as resampling, class weighting, and focal loss. However, these methods did not yield significant performance gains.

7 Discussion

Table 4 presents the Macro-F1 scores across six target languages for each emotion class, and Figure 1 shows the training distribution of those emotions. Together, they reveal several important insights about model performance and class imbalance.

7.1 Emotion Analysis

A clear pattern of class imbalance emerges across the training data, with *anger* being the most dominant emotion in nearly all languages—accounting for over 50% of the samples in languages such as Amharic, Igbo, and Swahili.

This skewness partially explains the relatively strong F1 scores for anger in some cases (e.g., 0.67 in Amharic, 0.62 in Hausa). However, frequency alone does not guarantee high performance: for instance, Igbo shows high anger frequency but yields a very low F1 score (0.16), likely due to a combination of limited training samples and linguistic complexity.

Conversely, *disgust*, *surprise*, and *fear* are consistently underrepresented—virtually absent in Sundanese—and correspondingly result in very low F1 scores (e.g., Fear is 0.00 in Sundanese and Yoruba; Surprise is 0.02 in Igbo). This highlights the vulnerability of emotion classification models to sparse training data, particularly for nuanced emotions.

Despite only moderate representation in most languages, *joy* demonstrates relatively strong performance, especially in Sundanese (F1 = 0.83) and Amharic (F1 = 0.77). This suggests that the model is better able to generalize joy-related patterns, potentially due to more consistent lexical cues or semantic clarity across languages.

Among the studied languages, Amharic and Hausa exhibit the most balanced class distributions and, correspondingly, perform best on several emotion categories. Amharic achieves the highest F1 scores for emotions such as *disgust*, *joy*, and *sadness*, with Hausa closely following. These results reaffirm the value of balanced training data in enhancing model robustness.

In contrast, languages with extreme class imbalance exhibit generally poor generalization, with uniformly low F1 scores across emotions—underscoring the limitations of even transfer learning in such settings.

Overall, this analysis reveals a strong link between class balance in training data and downstream performance, while also exposing persistent language-specific challenges. Addressing these issues may require more nuanced interventions, such as targeted data augmentation, synthetic oversampling of minority classes, or techniques like label smoothing to mitigate imbalance-induced bias in low-resource emotion detection.

7.2 Error Analysis

Despite AfroEmo's strong overall performance, several limitations persist in AfroEmo and require further investigation.

Emotion Confusion: The model frequently misclassifies *fear* as *sadness*, highlighting limitations in capturing fine-grained emotional distinctions. This suggests the need for more nuanced emotional representations, particularly for culturally sensitive emotions (see Table 4).

Sparse Classes: Emotions such as *disgust* and

surprise consistently yield low F1 scores, which correlates with their sparse presence in the training data. This imbalance constrains the model's learning capacity. Future work may benefit from class-balancing strategies such as data augmentation, oversampling, or curriculum learning.

Generalization: AfroEmo performs well on in-domain text, however, its performance deteriorates on unseen distributions. This underscores the need for more robust domain adaptation techniques to improve generalizability in real-world multilingual contexts.

7.3 Ablation Study

To quantify the contribution of each stage of the AfroEmo architecture, we conduct a series of ablation experiments focusing on adaptive pre-training, and fine-tuning for perceived emotions.

Removing the masked language modeling (MLM) step and directly fine-tuning Afro-XLM-R on the emotion dataset results in an average Macro-F1 drop of approximately 10% for low-resource languages like Hausa and Swahili. This demonstrates the critical role of domain-adaptive pre-training in enhancing contextual understanding of emotionally rich, morphologically complex languages.

Substituting Afro-XLM-R with a general-purpose multilingual model (XLM-RoBERTa) leads to consistent performance degradation across all metrics. This highlights the importance of Afro-XLM-R's linguistic specialization for African languages, which better captures regional nuances and syntactic patterns.

Excluding perceived emotion annotations and relying solely on explicit emotion labels causes a 4% decline in Macro-F1, particularly affecting culturally variable emotions like *fear* and *surprise*. This confirms the utility of incorporating perceived emotional signals to improve emotion disambiguation across culturally diverse language data.

To conclude, each component—adaptive pre-training, use of Afro-XLM-R, and perceived emotion integration—plays a vital role in enabling state-of-the-art performance in multilingual, low-resource emotion detection.

Conclusions

We present a novel approach to emotion detection in low-resource languages using an AfroXLM-R-based architecture-AfroEmo. It demonstrates strong performance, particularly on Amharic, setting a new benchmark for multilingual emotion classification. However, variation in performance across languages highlights the persistent challenges posed by limited training data and cultural variability in emotional expression. To mitigate class imbalance and further enhance accuracy, future work will explore integration of ensemble learning and data augmentation. We also plan to implement zero-shot learning for extremely low-resource settings. These directions aim to strengthen multilingual emotion detection and contribute to broader advancements in low-resource natural language understanding.

Acknowledgments

We sincerely thank the SemEval-2025 Task 11 organizers for designing an engaging shared task and for releasing a valuable multilingual, multi-label emotion classification dataset that spans diverse low-resource languages. We also extend our gratitude to the reviewers for their insightful comments, constructive feedback, and generous support, all of which significantly contributed to the improvement of this work.

Limitations

Despite the promising outcomes, our model has the following limitations.

Emotion Ambiguity: Certain emotions, particularly *Surprise* and *Fear*, are highly dependent on cultural context, making consistent classification challenging.

Generalization Challenges: The model's effectiveness diminishes on out-of-domain test sets, emphasizing the necessity of domain adaptation techniques.

References

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

Sulaiman Almutairi, Mohammed Abohashrh, Hasanain Hayder Razzaq, Muhammad Zulqarnain, Abdallah Namoun, and Faheem Khan. 2024. A hybrid deep learning model for predicting depression symptoms from large-scale textual dataset. *IEEE Access*.

Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and

Alexander Gelbukh. 2023. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534.

Babs. 2023. babs/hausa-scrapped-texts. https://huggingface.co/datasets/babs/hausa-scrapped-texts. Accessed: 2025-04-14.

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. Semeval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295.

Christos Baziotis, Athanasiou Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 245–255.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540.

Rafael A Calvo, Sidney D'Mello, Jonathan Matthew Gratch, and Arvid Kappas. 2015. *The Oxford handbook of affective computing*. Oxford University Press.

Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. 2017. Affective computing and sentiment analysis. *A practical guide to sentiment analysis*, pages 1–10.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 05, pages 7594–7601.

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.

Luna De Bruyne. 2023. The paradox of multilingual emotion detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 458–466.

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, NAACL-HLT*, volume 1, pages 4171–4186.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

D. Gurgurov. 2023a. Dgurgurov/igbo_sa. https://huggingface.co/datasets/DGurgurov/igbo_sa. Accessed: 2025-04-14.

D. Gurgurov. 2023b. Dgurgurov/sundanese_sa. https://huggingface.co/datasets/DGurgurov/sundanese_sa. Accessed: 2025-04-14.

D. Gurgurov. 2023c. Dgurgurov/swahili_sa. https://huggingface.co/datasets/DGurgurov/swahili_sa. Accessed: 2025-04-14.

Fasih Haider, Senja Pollak, Pierre Albert, and Saturnino Luz. 2021. Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods. *Computer Speech & Language*, 65:101119.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Prachi Juyal and Amit Kundalya. 2023. Emotion detection from text: Classification and prediction of moods in real-time streaming text. In *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 46–52.

KerVerse. 2023. Kerverse/amharic_stories. https://huggingface.co/datasets/KerVerse/Amharic_Stories. Accessed: 2025-04-14.

Sheetal D. Kusal, Shruti G. Patil, Jyoti Choudrie, and Ketan V. Kotecha. 2024. Understanding the performance of ai algorithms in text-based emotion detection for conversational agents. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(8).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rakshit R Malagi, Yogith R, Sai Prashanth T K, Ashwini Kodipalli, Trupthi Rao, and Rohini B R. 2023. Emotion detection from textual data using supervised machine learning models. In *2023 4th International Conference for Emerging Technology (INCET)*, pages 1–5.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M Mohammad, Sebastian Ruder,

et al. 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, et al. 2025a. 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, Shamsuddeen Hassan othersMuhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, et al. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, 7:100943–100953.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2024. Emomix-3l: A code-mixed dataset for bangla-english-hindi emotion detection. *arXiv arXiv:2405.06922*.

Kushal Tatariya, Heather Lent, and Miryam de Lhoneux. 2023. Transfer learning for code-mixed data: Do pretraining languages matter? In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 365–378, Toronto, Canada. Association for Computational Linguistics.

Harshita Sanjay Trimukhe, Rutuja Anil Pagare, Shreya Sandeep Salunke, Afeefa Rafeeque, Rasheed Noor, and Salman Baig. 2024. Comparison of emotion through text analysis using various deep learning models. In *2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, pages 1–8.

Chaithra I V and Samanvaya K J. 2024. Text-based emotion recognition using deep learning. In *2024 Second International Conference on Advances in Information Technology (ICAIT)*, volume 1, pages 1–7.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 470–475, Bangkok, Thailand. Association for Computational Linguistics.

Xulang Zhang, Rui Mao, and Erik Cambria. 2024. Multilingual emotion recognition: Discovering the variations of lexical semantics between languages. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE.