YNWA_PZ at SemEval-2025 Task 11: Multilingual Multi-Label Emotion Classification

Mohammad Sadegh Poulaei[1], Erfan Zare[1], Mohammad Reza Mohammadi[1], Sauleh Etemadi[2],

[1]Iran University of Science and Technology, [2]University of Birmingham,

m_poulaei@comp.iust.ac.ir, e_zare@elec.iust.ac.ir, mrmohammadi@iust.ac.ir, s.eetemadi@bham.ac.uk

Abstract

This paper investigates multilingual emotion classification across three tasks: binary classification, intensity estimation, and cross-lingual emotion detection. To address challenges posed by linguistic diversity and limited annotated data, we explore a range of deep learning approaches, including transformer-based embeddings and traditional classifiers. Following extensive experimentation, language-specific embedding models were selected as the final approach due to their superior capability to capture linguistic and cultural nuances. Evaluations on both high- and low-resource languages demonstrate that this method yields strong performance, achieving competitive macro-average F1 scores across tasks. Notably, in the cross-lingual detection task, our approach secured first-place rankings in Oromo, Tigrinya, and Kinyarwanda, driven by the integration of advanced preprocessing techniques and tailored language modeling. Despite these advances, challenges persist due to data scarcity in underrepresented languages and the inherent complexity of emotional expression. This study underscores the importance of developing robust, language-aware emotion recognition systems and highlights future directions, including the expansion of multilingual datasets and continued refinement of modeling techniques.

1 Introduction

The analysis and processing of emotions from textual data have become crucial in understanding human communication across different languages and cultures. This study focuses on the detection and classification of emotions across diverse linguistic contexts, spanning regions from South America to East Asia. Our objective is to categorize emotions into key dimensions, namely sadness, anger, fear, disgust, joy, and surprise, while considering cross-lingual variations and linguistic complexities.

To address these challenges, we structure our study into three distinct tracks: (1) Track A involves binary emotion classification, determining whether a given text expresses a particular emotion; (2) Track B measures the intensity of emotions on a scale from 0 to 3, enabling a more granular understanding of emotional expressions; and (3) Track C explores cross-lingual emotion detection, facilitating insights into emotional patterns across different languages.

Understanding emotions based on textual data plays a pivotal role in various applications, including social media analysis, behavioral research, and the study of emotions' influence on social interactions. Our work contributes to the development of robust emotion recognition systems, enabling better comprehension of multilingual emotional expressions and their implications in computational linguistics.

Despite the significant advancements in emotion classification, several challenges persist. Some languages exhibit highly complex grammatical structures, making it difficult to train effective models. Additionally, the classification of emotions in low-resource languages is hindered by data scarcity and syntactic intricacies. Furthermore, certain machine learning models demonstrate suboptimal performance when applied to multilingual emotion classification, necessitating the development of novel techniques to enhance model adaptability and generalization.

To address these limitations, we present a comprehensive analysis of state-of-the-art methodologies and evaluate their effectiveness across multiple languages. Our findings highlight the critical role of innovative preprocessing techniques, domain adaptation strategies, and transfer learning in improving multilingual emotion classification.

All code implementations, including the models and experimental setups employed in this study, are publicly available on GitHub:[1]. This repository pro-

[1]https://github.com/YNWA-PZ/SemEval2025-task11

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 508–521

vides full documentation of our methodologies, experimental results, and final model architectures.

2 Related Work

Multi-label emotion detection has emerged as a significant task in NLP[2], particularly for low-resource languages. The task is structured in two main output formats: (1) a binary format, which indicates whether an emotion is present in the text, and (2) an intensity scale ranging from 0 to 3, which represents the strength of the emotion in the text.

Given that this task follows a text classification paradigm, various models have been explored to identify the most effective architectures. A considerable amount of research has focused on evaluating different structures to determine the optimal approach. In (Wang et al., 2016), a combination of LSTM[3] networks and CNNs[4] was explored, where various model configurations were compared based on their F1-score performances. These insights were leveraged to identify suitable model structures for developing a custom model tailored to the specific requirements of this task.

For low-resource languages, text preprocessing plays a crucial role in improving model performance. The work presented in (Muhammad et al., 2023) highlighted the effectiveness of multiple preprocessing algorithms specifically designed for African languages. The study demonstrated that well-structured preprocessing pipelines lead to better text representations, ultimately improving classification accuracy.

Moreover, datasets specifically curated for emotion detection in underrepresented languages have been explored. The datasets presented in (Muhammad et al., 2025a) and (Belay et al., 2025) serve as essential resources for training models and evaluating performance in real-world settings. These datasets enable the training of robust models capable of handling linguistic diversity.

To enhance model performance, modifications to existing architectures have been proposed. Based on the insights from (Wang et al., 2016), additional layers were incorporated into custom models to improve the representation of low-resource languages. This ensures that the models can capture intricate linguistic patterns that might otherwise be overlooked.

[2]natural language processing
[3]Long Short-Term Memory
[4]Convolutional Neural Networks

3 System Overview

In this section, we present a comprehensive overview of our system for multi-label text classification, which integrates various deep learning architectures and machine learning classifiers. The system follows a pipeline that includes text preprocessing, feature extraction using neural network models, and classification through different machine learning algorithms.

3.1 Preprocessing

The preprocessing pipeline involves several steps to clean and standardize the text data. These include converting text to lowercase, removing unnecessary whitespace, filtering out special characters, URLs, and emojis by replacing with their textual description, normalizing tokens, performing language-specific tokenization, and removing stopwords. These steps ensure the data is consistent and suitable for NLP tasks.

3.2 Feature Extraction

To extract features, we employed a diverse range of models, including LSTM networks, MLMs[5], and LLMs[6]. The LLMs were fine-tuned using LoRA[7] (Hu et al., 2021), a parameter-efficient tuning method that facilitates task-specific adaptation while maintaining computational efficiency.

The extracted feature vectors were derived using two distinct approaches. The first approach utilized the output from the embedding layer of the models, which captures contextual word representations in a lower-dimensional vector space. The second approach involved extracting the final hidden state of the neural network, which encapsulates high-level semantic information of the text.

3.3 Classification Approach

Following feature extraction, we applied multiple classification algorithms to perform the multi-label classification task. One of the classifiers used was the MLP[8], a feedforward artificial neural network capable of modeling complex relationships between the extracted features and the target labels. Additionally, the system employed XGBoost, a gradient boosting framework renowned for its effectiveness in structured data classification (Chen

[5]Multilingual Language Models
[6]Large Language Models
[7]Low-Rank Adaptation
[8]Multi-Layer Perceptron

and Guestrin, 2016). Furthermore, SVMs were utilized as a classification method due to their ability to operate effectively in high-dimensional feature spaces by identifying optimal hyperplanes for classification (Cortes and Vapnik, 1995).

For languages with sufficient pretrained models available using MTEB[9](Muennighoff et al., 2022), we identified the best-performing embedding model and paired it mainly with SVM as the classifier. This approach leverages the strengths of the pretrained embedding models in capturing language-specific nuances, while the SVM classifier ensures robust performance for multi-label classification. On the other hand, for languages with limited pretrained resources, we utilized the multilingual embedding model "Multilingual E5 large instruct" (Wang et al., 2024) in combination with XGBoost as the classifier. The model, designed to generalize across diverse languages, enabled the system to maintain high performance even in resource-constrained settings.

4 Experimental Setup

This section outlines the experimental setup, including data splits, preprocessing, hyperparameter tuning, computational resources, and the tools and libraries used, aiming for reproducibility and transparency. All experiments and evaluation protocols in this work are conducted following the guidelines specified in SemEval-2025 Task 11 (Muhammad et al., 2025b), which establishes the framework for text-based emotion detection.

4.1 Data Splits and Usage

The dataset(Muhammad et al., 2025a) was divided into three subsets: training, development (validation), and testing. Specifically, 80% of the training dataset was allocated for training, while the remaining 20% was reserved for validation to facilitate model selection. Once the best-performing model was identified during the validation phase, the entire training and development datasets were combined to retrain the final model. This final model was then evaluated on the test dataset, which was held out during the entire training process to ensure an unbiased assessment of the model's generalization performance. This approach adheres to standard practices in machine learning research to prevent data leakage and ensure robust evaluation (Goodfellow et al., 2016).

[9]Massive Text Embedding Benchmark

4.2 Preprocessing

Preprocessing of the dataset was performed using the `clean-text` library. The preprocessing pipeline involved multiple steps to clean and standardize the text data. Initially, all text was converted to lowercase, and unnecessary whitespace was removed to eliminate redundancy. Special characters, URLs, and emojis were filtered out using regular expressions. Emojis were replaced by their corresponding textual descriptions (e.g., ☺→ "smiling face"). Punctuation was also removed, and tokenization was performed using language-specific tokenizers to ensure optimal segmentation, and stopwords were removed to further reduce noise. These steps ensured the data was clean and consistent across all subsets. Preprocessing was applied consistently to the training, validation, and test datasets to avoid introducing biases or inconsistencies. Such preprocessing steps have been shown to improve the performance of NLP models by reducing noise and simplifying the input representations (Zhang and Wang, 2020).

4.3 Hyperparameter Tuning

Hyperparameter tuning used Optuna (Akiba et al., 2019) to optimize SVM and XGBoost hyperparameters. Bayesian optimization balanced exploration and exploitation, with configurations assessed on the validation set. The best configuration was selected based on the performance metric.

4.4 Model Training and Optimization

The model fine-tuning with LoRA, and the training of the MLP and XGBoost models, utilized Binary Cross-Entropy (BCE) as the loss function for Tracks A and C, and Cross-Entropy for Track B, owing to its appropriateness for classification tasks. Meanwhile, the training of the SVM model employed hinge loss.Using LoRA, we fine-tuned the Q, K, and V matrices for feature extractor transformer models, as shown in Table 3. Given the unbalanced dataset, a weighted loss approach was employed to ensure that the model adequately learned from all classes. Optimization for fine-tuning deep learning models was performed using the `AdamW` optimizer, which improves upon the standard Adam optimizer by decoupling weight decay and learning rate updates (Loshchilov and Hutter, 2019). To further enhance training stability and convergence, a cosine annealing learning rate scheduler with restarts(Loshchilov and Hutter, 2017) was em-

Table 1: Results across Track A, B, and C showing macro-average F1 scores of Our Model , Paraticipants Best Model scores, Task Dataset Best Model with Baseline(Muhammad et al., 2025a) and rankings.

		Track A				Track B					Track C				
Language	**Our Model**	**Ours**	**BP***	**Base**	**Rank**	**Ours**	**BP***	**BDP****	**Base**	**Rank**	**Ours**	**BP***	**BDP****	**Base**	**Rank**
Afrikaans(afr)	(Wang et al., 2024) + SVM	54.01	69.86	37.14	13/32	—	—	—	—	—	54.01	70.50	61.28	35.04	4/12
Amharic(amh)	(Benmounah et al., 2023) + SVM	61.20	77.31	63.83	18/40	49.42	85.58	—	50.79	12/20	61.20	66.68	—	48.66	4/11
Algerian Arabic(arq)	(Wang et al., 2024) + SVM	51.07	66.87	41.41	18/36	36.54	64.97	36.37	1.64	15/23	51.07	58.75	55.75	33.78	4/12
Moroccan Arabic(ary)	(Wang et al., 2024) + SVM	51.88	62.92	47.16	17/35	—	—	—	—	—	51.88	63.22	52.76	35.46	4/10
Chinese(chn)	(iampanda, 2024) + SVM	56.65	70.94	53.08	25/36	48.47	72.24	51.86	40.53	15/24	56.65	68.89	55.23	24.56	5/12
German(deu)	(Wang et al., 2024) + SVM	60.60	73.99	64.23	21/44	54.10	76.57	56.21	56.21	15/24	60.60	72.67	59.17	46.84	4/12
English(eng)	(Zhang et al., 2025) + SVM	73.97	82.30	70.83	28/74	68.81	84.04	64.15	64.15	20/36	73.97	79.69	65.58	37.54	3/12
Spanish(esp)	(Wang et al., 2024) + SVM	76.19	84.88	77.44	24/44	66.70	80.80	72.59	72.59	20/26	76.19	83.11	73.29	57.37	3/13
Hausa(hau)	(Dobler and de Melo, 2023) + SVM	63.22	75.07	59.55	16/36	58.42	77.00	39.16	27.03	12/23	63.22	70.88	51.91	31.98	**2/11**
Hindi(hin)	(Wang et al., 2024) + SVM	80.32	92.57	85.51	30/39	—	—	—	—	—	80.32	91.87	79.73	13.75	4/14
Igbo(ibo)	(Wang et al., 2024) + SVM	50.93	60.01	47.90	11/30	—	—	—	—	—	50.93	60.47	37.40	7.49	**2/9**
Indonesian(ind)	(Wang et al., 2024) + XGB	—	—	—	—	—	—	—	—	—	35.64	67.24	57.29	37.64	13/15
Javanese(jav)	(Wang et al., 2024) + XGB	—	—	—	—	—	—	—	—	—	25.62	46.38	50.47	46.38	10/11
Kinyarwanda(kin)	(Wang et al., 2024) + SVM	51.94	65.74	46.29	5/28	—	—	—	—	—	**51.94**	**51.94**	34.36	18.38	**1/8**
Marathi(mar)	(Wang et al., 2024) + SVM	81.10	88.43	82.20	21/37	—	—	—	—	—	81.10	90.29	77.24	77.24	4/11
Oromo(orm)	(Wang et al., 2024) + SVM	54.31	61.64	12.63	9/31	—	—	—	—	—	**54.31**	**54.31**	—	26.17	**1/9**
Nigerian-Pidgin(pcm)	(Wang et al., 2024) + SVM	53.09	67.40	55.50	19/30	—	—	—	—	—	53.09	67.40	48.67	1.01	3/8
Pt*** Brazilian(ptbr)	(Souza et al., 2020) + SVM	47.99	68.33	42.57	23/37	38.20	71.00	46.72	29.74	19/23	47.99	62.91	51.60	41.84	5/11
Pt*** Mozambican(ptmz)	(Wang et al., 2024) + SVM	50.08	54.77	45.91	5/32	—	—	—	—	—	50.08	55.54	40.44	29.67	**2/11**
Romanian(ron)	(Wang et al., 2024) + SVM	73.75	79.43	76.23	14/39	57.61	72.60	57.69	55.66	14/22	73.75	76.70	76.23	76.23	4/13
Russian(rus)	(Snegirev et al., 2025) + SVM	82.42	90.08	83.77	28/44	78.41	92.54	87.66	87.66	18/25	82.42	90.58	76.97	70.43	4/14
Somali(som)	(Wang et al., 2024) + SVM	48.26	57.65	45.93	7/29	—	—	—	—	—	48.26	47.79	—	27.27	3/10
Sundanese(sun)	(Wang et al., 2024) + SVM	42.48	54.97	37.31	17/32	—	—	—	—	—	42.48	46.66	46.33	19.43	3/9
Swahili(swa)	(Wang et al., 2024) + SVM	29.52	38.56	22.65	13/29	—	—	—	—	—	29.52	38.05	33.27	18.99	3/11
Swedish(swe)	(Wang et al., 2024) + SVM	56.51	62.62	51.98	12/34	—	—	—	—	—	56.51	64.53	51.18	51.18	4/11
Tatar(tat)	(Wang et al., 2024) + SVM	64.32	84.59	53.94	15/31	—	—	—	—	—	64.32	78.86	60.66	44.54	3/9
Tigrinya(tir)	(Wang et al., 2024) + SVM	52.37	59.05	46.28	6/28	—	—	—	—	—	**52.37**	**52.37**	—	33.93	**1/8**
Ukrainian(ukr)	(Sturua et al., 2024) + SVM	48.62	72.56	53.45	26/36	42.55	70.75	43.54	39.94	13/21	48.62	70.18	54.76	49.56	9/15
Emakhuwa(vmw)	(Sturua et al., 2024) + SVM	16.81	32.50	12.14	11/20	—	—	—	—	—	16.80	21.04	20.41	5.22	4/7
isiXhosa(xho)	(Wang et al., 2024) + XGB	—	—	—	—	—	—	—	—	—	16.64	44.26	30.79	12.73	4/8
Yoruba(yor)	(Wang et al., 2024) + SVM	34.09	46.13	9.22	7/30	—	—	—	—	—	34.09	35.95	27.44	5.33	3/8
isiZulu(zul)	(Wang et al., 2024) + XGB	—	—	—	—	—	—	—	—	—	16.35	39.69	22.03	15.26	6/9

BP=result of rank 1*
*BDP**=best result of dataset paper(Muhammad et al., 2025a)*
*pt***=Portuguese*
The Best result of dataset paper for Track A is identical to that of Track C.

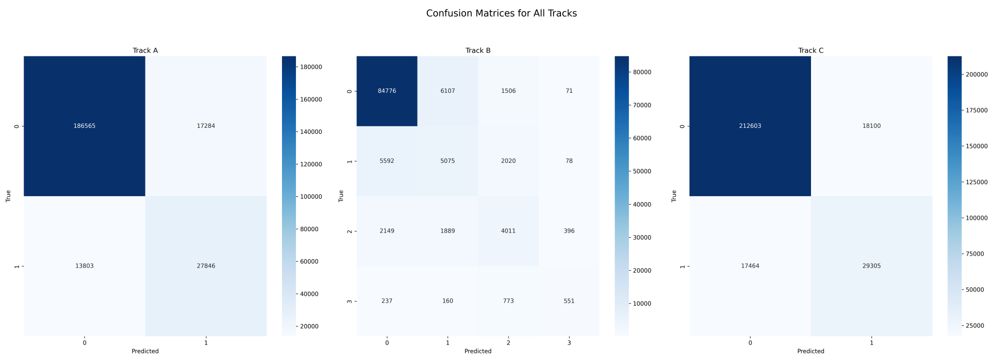

Figure 1: Confusion matrices for each language across Tracks A, B, and C.

ployed. This scheduling approach helped adaptively reduce the learning rate over time, facilitating better exploration of the loss landscape and improving generalization. The model was trained for a fixed number of epochs, and early stopping was used to terminate training if the validation performance plateaued, thus avoiding overfitting.

4.5 Tools and Libraries

The implementation of the experiments utilized several state-of-the-art tools and libraries. The deep learning models were implemented and trained using `PyTorch` (Paszke et al., 2019). For data manipulation and evaluation metrics, `Scikit-Learn` was employed (Pedregosa et al., 2011). Gradient boosting models were benchmarked using `XGBoost` (Chen and Guestrin, 2016). Pre-trained transformer models were fine-tuned using `Hugging Face Transformers` (Wolf et al., 2020). Additionally, the `Sentence Transformers` library was used to load embedding models (Reimers and Gurevych, 2019)(Reimers and Gurevych, 2020a). These tools and libraries are well-regarded in the machine learning community and were chosen for their reliability and performance.

4.6 Computational Resources

Experiments used a Kaggle Tesla P100 GPU for efficient model training, evaluation, and hyperparameter tuning, ensuring reproducibility with com-

Text	Type	Anger	Fear	Joy	Sadness	Surprise
I'm just numb.	Truth	0	0	0	1	0
	Pred	0	1	0	1	0
At the time it didn't seem to bother me.	Truth	0	0	0	1	0
	Pred	0	0	0	0	0
I found out six weeks before the wedding that my dad had only six weeks to live (he had cancer for two years... a fact she was fully aware of).	Truth	1	1	0	1	1
	Pred	0	1	0	1	0

Table 2: Error Analysis Table of language English for track A

parable hardware.

5 Results

Extensive experiments were conducted on multiple models to determine the most effective approach for multi-label emotion detection across various languages. The selected model was trained on datasets corresponding to each language, and its performance was analyzed using the test dataset. The evaluation results are presented in Table 1. More detailed results and additional analysis can be found in the Appendix A.

Notably, our approach achieved first rank in the Oromo, Tigrinya, and Kinyarwanda languages in Track-C of the competition. This strong performance highlights the effectiveness of the use of language-specific model embeddings tailored to the linguistic characteristics of each language.

A comparison of our findings with reference studies (Muhammad et al., 2025a) highlights the effectiveness of our approach. By leveraging domain-specific model embeddings, our models were able to bridge the gap in emotion classification for low-resource languages.

Table 2 highlights key limitations in the model's contextual understanding. For instance, the model misidentified fear with sadness in "I'm just numb," due to an oversimplified link between numbness and fear, showing lexical misinterpretation without context. Another example shows the model's failure to recognize temporal contrast in "at the time," missing the current sadness implied by past indifference, indicating a need for deeper semantic processing. In the third example, the model detected sadness and fear in a father's terminal illness revelation but missed anger and surprise embedded contextually, particularly the implicit anger towards "she" who knew about the cancer and the surprise of receiving life-altering news before a significant event, revealing deficiencies in extracting emotional implications from complex narratives.

Figure 1 presents the pooled confusion matrices for Tracks A, B, and C, highlighting the classification performance and misclassifications across different intensity levels and languages.

6 Conclusion

This study presented a comprehensive examination of multilingual multi-label emotion detection, addressing binary classification, intensity estimation, and cross-lingual detection tasks. Our findings indicate that language-specific embedding models, when paired with classifiers such as SVM and XGBoost, offer a robust approach to capturing the nuanced linguistic and cultural features inherent in diverse textual data. The experimental results, measured in competitive macro-average F1 scores, underscore the potential of these tailored models to bridge performance gaps, particularly in low-resource languages where data scarcity and complex grammatical structures present significant challenges.

The significance of this research lies in its demonstration that integrating innovative preprocessing techniques with state-of-the-art embedding models can lead to substantial improvements in emotion recognition performance. This has broad implications for applications in social media analysis, behavioral research, and other domains where understanding nuanced emotional expressions is crucial.

Nonetheless, current limitations in multilingual emotion analysis include the lack of annotated data for underrepresented languages and challenges in capturing nuanced emotional expressions, both of which hinder model performance. Future research should prioritize expanding multilingual datasets, improving preprocessing techniques, and developing new architectures to boost model generalization and adaptability. Fine-tuning models for low-resource languages could also enhance emotion detection accuracy, advancing the field and creating more effective, language-aware emotion recognition systems.

References

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. Dziribert: a pre-trained language model for the algerian dialect. *arXiv preprint arXiv:2109.12346*.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

David Adelani. 2023a. bert-base-multilingual-cased-finetuned-kinyarwanda. `https://huggingface.co/Davlan/bert-base-multilingual-cased-finetuned-kinyarwanda`.

David Adelani. 2023b. xlm-roberta-large-finetuned-kinyarwanda. `https://huggingface.co/Davlan/xlm-roberta-large-finetuned-kinyarwanda`.

Meta AI. 2025. Llama 3.2 1b instruct model. `https://github.com/meta/llama`. A fine-tuned version of the Llama 3.2 model optimized for instruction-following tasks.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

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, Abu Dhabi, UAE. Association for Computational Linguistics.

Zakaria Benmounah, Abdennour Boulesnane, Abdeladim Fadheli, and Mustapha Khial. 2023. Sentiment analysis on algerian dialect with transformers. *Applied Sciences*, 13(20):11157.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. *Preprint*, arXiv:2010.10906.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Aaron Chibb. 2023. German_semantic_sts_v2. `https://huggingface.co/aari1995/German_Semantic_STS_V2`.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Zihang Dai, Guokun Lai, Yiming Yang, and Quoc V. Le. 2020. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. *Preprint*, arXiv:2006.03236.

Davlan. 2025. Bert base multilingual cased fine-tuned on amharic. `https://huggingface.co/Davlan/bert-base-multilingual-cased-finetuned-amharic`. Accessed: 2025-02-24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Konstantin Dobler and Gerard de Melo. 2023. Focus: Effective embedding initialization for monolingual specialization of multilingual models. *Preprint*, arXiv:2305.14481.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Ricardo Filho. 2023. bert-base-portuguese-cased-nli-assin-2. `https://huggingface.co/ricardo-filho/bert-base-portuguese-cased-nli-assin-2`.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2023. Darijabert: a step forward in nlp for the written moroccan dialect.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Daniel Heinz. 2023. e5-base-sts-en-de: A bilingual text embedding model for english and german. Hugging Face Model Hub. Accessed: 2025-02-25.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

iampanda. 2024. zpoint-large-embedding-zh. https://huggingface.co/iampanda/zpoint_large_embedding_zh.

Ananya Joshi, Aditi Kajale, Janhavi Gadre, Samruddhi Deode, and Raviraj Joshi. 2022. L3cube-mahasbert and hindsbert: Sentence bert models and benchmarking bert sentence representations for hindi and marathi. *arXiv preprint arXiv:2211.11187*.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Yurii Laba, Volodymyr Mudryi, Dmytro Chaplynskyi, Mariana Romanyshyn, and Oles Dobosevych. 2023. Contextual embeddings for Ukrainian: A large language model approach to word sense disambiguation. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 11–19, Dubrovnik, Croatia. Association for Computational Linguistics.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Shiyu Li, Yang Tang, Shizhe Chen, and Xi Chen. 2024. Conan-embedding: General text embedding with more and better negative samples. *Preprint*, arXiv:2408.15710.

lier007. 2023. xiaobu-embedding-v2: A text embedding model. Hugging Face Model Hub. Accessed: 2025-02-25.

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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Sgdr: Stochastic gradient descent with warm restarts. *Preprint*, arXiv:1608.03983.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *International Conference on Learning Representations*.

Rui Melo. 2023. bert-large-portuguese-cased-sts. https://huggingface.co/rufimelo/bert-large-portuguese-cased-sts.

Benjamin Minixhofer. 2023. roberta-large-wechsel-ukrainian. https://huggingface.co/benjamin/roberta-large-wechsel-ukrainian.

Isabelle Mohr, Markus Krimmel, Saba Sturua, Mohammad Kalim Akram, Andreas Koukounas, Michael Günther, Georgios Mastrapas, Vinit Ravishankar, Joan Fontanals Martínez, Feng Wang, and 1 others. 2024. Multi-task contrastive learning for 8192-token bilingual text embeddings. *arXiv preprint arXiv:2402.17016*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ayele, Saif M Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023. Semeval-2023 task 12: sentiment analysis for african languages (afrisenti-semeval). *arXiv preprint arXiv:2304.06845*.

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, and 29 others. 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, 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, and 2 others. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *Preprint*, arXiv:2108.08877.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *Preprint*, arXiv:1904.08375.

Finbarrs Oketunji. 2024a. pmmlv2-fine-tuned-hausa. https://huggingface.co/0xnu/pmmlv2-fine-tuned-hausa.

Finbarrs Oketunji. 2024b. pmmlv2-fine-tuned-igbo. https://huggingface.co/0xnu/pmmlv2-fine-tuned-igbo.

Serry Taiseer Sibaee Omer Nacar, Anis Koubaa and Lahouari Ghouti. 2025. Gate: General arabic text embedding for enhanced semantic textual similarity with hybrid loss training. Submitted to COLING 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rasyosef. 2025a. Llama 3.2 1b fine-tuned on amharic. https://huggingface.co/rasyosef/Llama-3.2-1B-Amharic. Accessed: 2025-02-24.

Rasyosef. 2025b. Roberta amharic text embedding (medium). https://huggingface.co/rasyosef/roberta-amharic-text-embedding-medium. Accessed: 2025-02-24.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020a. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020b. Making monolingual sentence embeddings multilingual using knowledge distillation. *Preprint*, arXiv:2004.09813.

Manuel Romero. 2023. multilingual-e5-large-ft-sts-spanish-matryoshka-768-64-5e. https://huggingface.co/mrm8488/multilingual-e5-large-ft-sts-spanish-matryoshka-768-64-5e.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Stefan Schweter. 2020. electra-base-ukrainian-cased-discriminator. https://huggingface.co/lang-uk/electra-base-ukrainian-cased-discriminator.

sentence transformers. 2024. all-minilm-l12-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2.

Artem Snegirev, Maria Tikhonova, Anna Maksimova, Alena Fenogenova, and Alexander Abramov. 2025. The russian-focused embedders' exploration: rumteb benchmark and russian embedding model design. *Preprint*, arXiv:2408.12503.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

Akshita Sukhlecha. 2024. Bhasha-embed-v0. Hugging Face.

Gemma Team. 2024a. Gemma.

Qwen Team. 2024b. Qwen2.5: A party of foundation models.

Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional cnn-lstm model. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 225–230.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Rasy Yosef. 2025. Bert amharic text embedding (medium). `https://huggingface.co/rasyosef/bert-amharic-text-embedding-medium`. Accessed: 2025-02-24.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. Jasper and stella: distillation of sota embedding models. *Preprint*, arXiv:2412.19048.

Li Zhang and Jun Wang. 2020. A comprehensive guide to text data preprocessing. *Journal of Data Science*, 12:45–67.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. Gme: Improving universal multimodal retrieval by multimodal llms. *Preprint*, arXiv:2412.16855.

A Model Selection

The training dataset of Track A was utilized for model selection, where it was split into an 80/20 ratio to create a training set and a validation set. Several models were evaluated on this split to identify the optimal model based on performance metrics such as recall, precision, and F1-score. The model that achieved the best balance across these metrics was selected as the final model and applied consistently across all three tracks—A, B, and C. A detailed comparative analysis of the performance of different models for each language is presented in Table 3, where each row corresponds to a specific language and includes results for multiple models, with columns reporting recall, precision, and F1-score for each.

Table 3 provides a comprehensive comparison of the models' performance for each language, aiding in the selection of the final model applied to Tracks A, B, and C.

Table 3: Performance comparison of models for different languages

Language	Model	Metrics		
		Recall	Precision	F1-Score
All	(Radford et al., 2019) + MLP	28.42	71.37	39.98
	(AI, 2025) + MLP with LoRA	30.16	74.76	42.03
	(Abdin et al., 2024) + MLP with LoRA	39.69	76.57	51.83
	(Team, 2024b) + MLP with LoRA	33.02	73.59	44.43
	(Team, 2024a) + MLP with LoRA	21.80	61.13	30.95
	(Conneau et al., 2019) + MLP	64.51	50.24	56.38
	(Lewis et al., 2019) + MLP	32.88	69.08	43.60
	(Devlin et al., 2018) + MLP	35.63	71.75	46.82
	(Dai et al., 2020) + MLP	16.79	69.00	25.09
	(Clark et al., 2020) + MLP	26.58	69.27	37.40
Afrikaans(afr)	(Feng et al., 2022) + SVM	24.09	38.89	29.12
	(sentence transformers, 2024) + XGB	8.00	29.15	9.29
	(Wang et al., 2024) + SVM	58.10	49.18	52.19
	(Zhang et al., 2025) + XGB	12.55	27.55	15.15
	(Lee et al., 2024) + XGB	11.31	36.40	15.19
Amharic(amh)	(Yosef, 2025) + SVM	49.39	71.28	51.87
	(Davlan, 2025) + SVM	40.62	40.14	39.24
	(Wang et al., 2024) + SVM	64.80	56.98	59.97
	(Rasyosef, 2025a) + SVM	58.67	56.88	57.32
	(Rasyosef, 2025b) + XGB	35.33	45.68	39.67
	(Rasyosef, 2025b) + SVM	46.13	56.56	47.97
	(sentence transformers, 2024) + XGB	24.68	37.22	20.92
	(Sturua et al., 2024) + XGB	59.07	49.90	53.34
Algerian Arabic(arq)	(Benmounah et al., 2023) + SVM	46.68	55.78	50.33
	(Abdaoui et al., 2021) + SVM	47.16	53.15	49.59
	(Abdaoui et al., 2021) Sentiment + SVM	49.29	51.91	49.94
	(Wang et al., 2024) + SVM	41.80	63.27	48.48
	(Omer Nacar and Ghouti, 2025) + SVM	38.50	53.48	43.99
	(sentence transformers, 2024) + XGB	34.97	23.65	28.13
	(Sturua et al., 2024) + XGB	39.34	45.27	41.26
Moroccan Arabic(ary)	(Safaya et al., 2020) + SVM	52.82	53.64	51.81
	(Gaanoun et al., 2023) + SVM	37.12	58.05	41.10
	(Wang et al., 2024) + SVM	55.37	49.25	51.17
	(Omer Nacar and Ghouti, 2025) + SVM	33.34	52.32	39.49
	(sentence transformers, 2024) + XGB	28.19	20.76	22.07
	(Sturua et al., 2024) + XGB	36.60	50.29	40.24
Chinese(chn)	(Li et al., 2024) + DT	52.12	33.51	39.82
	(Li et al., 2024) + XGB	37.03	60.88	42.27
	(Li et al., 2024) + SVM	57.89	46.56	51.19
	(Wang et al., 2024) + SVM	54.94	43.62	48.25
	(Zhang et al., 2024) + DT	33.84	20.33	23.58
	(Zhang et al., 2024) + RF	7.22	24.03	9.44
	(Zhang et al., 2024) + XGB	12.77	22.87	15.90

Continued from previous page

Language	Model	Metrics		
		Recall	Precision	F1-Score
	(Zhang et al., 2024) + SVM	27.94	34.20	29.87
	(lier007, 2023) + XGB	33.09	70.84	37.98
	(lier007, 2023) + SVM	60.44	53.49	55.72
	(iampanda, 2024) + XGB	36.48	73.26	41.85
	(iampanda, 2024) + SVM	64.43	54.54	58.58
	(sentence transformers, 2024) + XGB	23.42	29.65	22.95
	(Sturua et al., 2024) + XGB	47.18	53.49	49.39
German(deu)	(sentence transformers, 2024) + XGB	16.77	45.99	23.26
	(Heinz, 2023) + SVM	41.64	61.08	45.45
	(Wang et al., 2024) + SVM	60.42	56.09	57.93
	(Chan et al., 2020) + XGB	33.69	60.54	40.14
	(Chibb, 2023) + SVM	56.87	57.48	56.25
	(Mohr et al., 2024) deu + XGB	36.24	58.28	42.87
	(Mohr et al., 2024) deu + SVM	45.68	60.41	50.25
	(Sturua et al., 2024) + XGB	35.78	55.25	42.43
	(Sturua et al., 2024) + SVM	55.39	59.72	56.63
	(Ni et al., 2021) + XGB	40.20	79.24	47.28
	(Wang et al., 2023) + XGB	38.57	73.09	46.18
English(eng)	(sentence transformers, 2024) + XGB	43.71	65.10	50.51
	(Devlin et al., 2018) embedding + XGB	30.60	50.23	35.01
	(Devlin et al., 2018) last hidden state + XGB	40.60	75.32	48.88
	(Wang et al., 2024) + SVM	76.79	70.85	73.43
	(Zhang et al., 2025) + XGB	60.58	80.60	68.30
	(Zhang et al., 2025) + XGB without preprocess	58.07	79.05	65.28
	(Zhang et al., 2025) + SVM	71.76	73.13	72.40
	(Liu et al., 2019) embedding + XGB	30.35	48.17	33.64
	(Liu et al., 2019) last hidden state + XGB	32.99	66.63	39.81
	(Ni et al., 2021) + XGB	59.06	74.62	64.62
	(Conneau et al., 2019) embedding + MLP	100	37.32	52.76
	(Conneau et al., 2019) embedding + Conv1D + MLP	55.96	25.96	34.93
	(Conneau et al., 2019) embedding + XGB	26.25	45.20	28.82
	(Conneau et al., 2019) last hidden state + XGB	25.36	52.99	28.90
	(Conneau et al., 2019) last hidden state + MLP	38.59	23.68	29.35
	(Lee et al., 2024) + XGB	54.59	80.89	61.75
	(Zhang et al., 2025) Stella + XGB	57.24	78.88	65.25
Spanish(esp)	(Cañete et al., 2020) + SVM	67.01	76.56	71.27
	(Mohr et al., 2024) es + SVM	75.78	82.24	78.46
	(Sturua et al., 2024) + SVM	79.36	78.13	78.64

Continued from previous page

Language	Model	Metrics		
		Recall	Precision	F1-Score
	(Sturua et al., 2024) + XGB	73.29	72.48	72.64
	(Romero, 2023) + SVM	74.98	79.15	76.68
	(sentence transformers, 2024) + XGB	43.83	59.21	49.60
Hausa(hau)	(Wang et al., 2024) + SVM	62.82	60.21	61.23
	(Sturua et al., 2024) + SVM	53.64	53.51	53.31
	(Sturua et al., 2024) + XGB	30.96	47.87	36.64
	(Oketunji, 2024a) + SVM	28.36	48.10	34.48
	(Dobler and de Melo, 2023) + SVM	65.98	60.71	62.79
	(sentence transformers, 2024) + XGB	19.64	40.70	25.44
Hindi(hin)	(Sukhlecha, 2024) + SVM	84.76	75.86	79.86
	(Wang et al., 2024) + SVM	83.90	78.58	80.77
	(Joshi et al., 2022) + SVM	74.14	80.15	76.83
	(Nogueira et al., 2019) + SVM	76.24	65.05	70.09
	(Feng et al., 2020) hin + SVM	72.03	81.16	76.02
	(sentence transformers, 2024) + XGB	18.86	30.58	20.81
	(Sturua et al., 2024) + XGB	74.66	73.77	73.96
Igbo(ibo)	(Feng et al., 2022) + SVM	42.41	51.86	44.88
	(Wang et al., 2024) + SVM	51.43	53.26	51.81
	(Oketunji, 2024b) + SVM	13.48	38.71	15.08
	(sentence transformers, 2024) + XGB	21.18	55.54	29.02
	(Sturua et al., 2024) + XGB	21.54	42.28	27.76
Kinyarwanda(kin)	(Feng et al., 2022) + SVM	39.35	51.88	42.27
	(Adelani, 2023a) + SVM	46.88	45.94	46.13
	(Wang et al., 2024) + SVM	50.97	45.98	48.09
	(Adelani, 2023b) + SVM	21.04	34.33	21.14
	(sentence transformers, 2024) + XGB	8.66	21.13	11.41
	(Sturua et al., 2024) + XGB	13.97	30.33	16.57
Marathi(mar)	(Wang et al., 2024) + SVM	79.03	80.20	79.38
	(Feng et al., 2022) + XGB	62.87	73.01	66.69
	(Feng et al., 2022) + SVM	76.97	76.80	76.81
	(sentence transformers, 2024) + XGB	23.70	44.08	27.77
	(Sturua et al., 2024) + XGB	71.47	68.67	69.62
Oromo(orm)	(Wang et al., 2024) + SVM	54.73	48.44	50.85
	(Feng et al., 2022) + XGB	20.67	26.50	20.60
	(Feng et al., 2022) + SVM	29.11	35.40	28.33
	(sentence transformers, 2024) + XGB	13.96	24.92	16.46
	(Sturua et al., 2024) + XGB	17.80	37.37	21.53
Nigerian-Pidgin(pcm)	(Wang et al., 2024) + SVM	52.03	49.58	50.24
	(Feng et al., 2022) + XGB	32.95	48.75	38.46
	(Feng et al., 2022) + SVM	42.93	49.03	44.81
	(sentence transformers, 2024) + XGB	33.40	38.45	33.22
	(Sturua et al., 2024) + XGB	39.59	45.08	41.18

Continued from previous page

Language	Model	Metrics		
		Recall	Precision	F1-Score
Pt* Brazilian(ptbr)	(Wang et al., 2024) + SVM	54.35	46.20	49.19
	(Filho, 2023) + SVM	43.54	44.40	42.79
	(Souza et al., 2020) + SVM	57.73	46.69	51.16
	(Melo, 2023) + SVM	36.33	56.90	38.76
	(Sturua et al., 2024) + SVM	40.74	47.26	42.83
	(Sturua et al., 2024) + XGB	49.22	45.96	42.94
	(sentence transformers, 2024) + XGB	14.17	32.08	18.89
Pt* Mozambican(ptmz)	(Wang et al., 2024) + SVM	54.53	45.70	48.21
	(Filho, 2023) + SVM	30.37	42.09	34.34
	(Souza et al., 2020) + SVM	41.23	41.07	39.80
	(Melo, 2023) + SVM	26.90	65.68	33.87
	(Sturua et al., 2024) + SVM	29.24	60.49	36.56
	(Sturua et al., 2024) + XGB	28.39	35.93	31.02
	(sentence transformers, 2024) + XGB	13.81	24.71	14.95
Romanian(ron)	(Wang et al., 2024) + SVM	75.68	71.90	72.99
	(Sturua et al., 2024) + SVM	70.59	68.50	69.43
	(Sturua et al., 2024) + XGB	50.75	72.67	57.49
	(Feng et al., 2022) + XGB	39.70	73.06	48.42
	(Feng et al., 2022) + SVM	59.75	72.83	64.04
	(sentence transformers, 2024) + XGB	37.86	51.06	41.81
Russian(rus)	(sentence transformers, 2024) + XGB	18.78	70.74	29.17
	(Wang et al., 2024) + SVM	77.26	73.24	75.11
	(Snegirev et al., 2025) + XGB	65.38	88.64	74.54
	(Snegirev et al., 2025) + SVM	79.32	84.12	81.57
	(Sturua et al., 2024) + XGB	67.59	61.99	64.03
Somali(som)	(Wang et al., 2024) + SVM	51.81	41.12	45.63
	(Feng et al., 2022) + XGB	29.44	37.43	32.13
	(Feng et al., 2022) + SVM	38.57	40.38	39.06
	(sentence transformers, 2024) + XGB	10.78	31.56	14.29
	(Sturua et al., 2024) + XGB	12.85	32.13	16.79
Sundanese(sun)	(Wang et al., 2024) + SVM	37.20	59.45	40.42
	(Feng et al., 2022) + XGB	23.84	41.55	29.44
	(Feng et al., 2022) + SVM	30.34	48.67	35.38
	(sentence transformers, 2024) + XGB	16.29	28.00	20.30
	(Sturua et al., 2024) + XGB	24.29	37.98	28.29
Swahili(swa)	(Wang et al., 2023) + XGB	24.90	26.12	25.30
	(Wang et al., 2024) + SVM	33.20	30.28	31.46
	(Feng et al., 2022) + XGB	21.21	22.85	21.21
	(Feng et al., 2022) + SVM	25.37	23.58	24.12
	(sentence transformers, 2024) + XGB	8.61	20.56	11.94
	(Sturua et al., 2024) + XGB	15.04	25.42	18.12
Swedish(swe)	(Wang et al., 2024) + XGB	32.53	43.31	35.73
	(Wang et al., 2024) + SVM	61.50	58.58	57.09
	(Kummervold et al., 2021) + XGB	30.61	48.66	35.00
	(sentence transformers, 2024) + XGB	18.63	25.11	19.29

Continued from previous page

Language	Model	Metrics		
		Recall	Precision	F1-Score
	(Sturua et al., 2024) + XGB	39.78	34.77	36.76
Tatar(tat)	(Wang et al., 2024) + SVM	50.32	60.76	54.40
	(Feng et al., 2022) + XGB	25.28	68.75	31.93
	(Feng et al., 2022) + SVM	39.81	61.07	46.44
	(sentence transformers, 2024) + XGB	4.44	27.17	7.26
	(Sturua et al., 2024) + XGB	19.88	36.94	25.43
Tigrinya(tir)	(Wang et al., 2024) + SVM	48.74	46.89	47.21
	(Feng et al., 2022) + XGB	23.31	32.37	24.96
	(Feng et al., 2022) + SVM	35.20	40.83	36.08
	(sentence transformers, 2024) + XGB	23.64	29.33	16.32
	(Sturua et al., 2024) + XGB	26.86	43.97	29.71
Ukrainian(ukr)	(Wang et al., 2024) + SVM	54.74	42.52	47.65
	(Schweter, 2020) + SVM	24.75	25.45	24.38
	(Snegirev et al., 2025) + SVM	39.01	74.30	45.40
	(Sturua et al., 2024) + SVM	45.92	56.02	49.43
	(Sturua et al., 2024) + XGB	60.39	37.82	45.43
	(Laba et al., 2023) + SVM	41.45	45.97	42.80
	(Minixhofer, 2023) + SVM	15.29	33.31	17.34
	(sentence transformers, 2024) + XGB	4.70	12.22	6.74
Emakhuwa(vmw)	(Wang et al., 2024) + SVM	14.43	22.04	15.46
	(Feng et al., 2022) + XGB	1.78	10.55	2.98
	(Feng et al., 2022) + SVM	5.81	21.38	8.79
	(sentence transformers, 2024) + XGB	3.35	20.55	5.63
	(Sturua et al., 2024) + XGB	1.35	7.56	2.27
Yoruba(yor)	(Feng et al., 2022) + SVM	20.88	38.07	25.91
	(Wang et al., 2024) + SVM	38.54	30.98	33.86
	(Reimers and Gurevych, 2020b) + SVM	37.51	28.02	28.31
	(Feng et al., 2022) + XGB	14.96	35.93	17.98
	(Feng et al., 2022) + SVM	19.34	37.22	22.82
	(sentence transformers, 2024) + XGB	9.49	20.80	11.34
	(Sturua et al., 2024) + XGB	9.58	22.61	9.81