PromotionGo at SemEval-2025 Task 11: A Feature-Centric Framework for Cross-Lingual Multi-Emotion Detection in Short Texts

Ziyi Huang
Hubei University
ziyihuang@hubu.edu.cn

Xia Cui
Manchester Metropolitan University
x.cui@mmu.ac.uk

Abstract

This paper presents our system for SemEval 2025 Task 11: Bridging the Gap in Text-Based Emotion Detection (Track A) (Muhammad et al., 2025b), which focuses on multi-label emotion detection in short texts. We propose a feature-centric framework that dynamically adapts document representations and learning algorithms to optimize language-specific performance. Our study evaluates three key components: document representation, dimensionality reduction, and model training in 28 languages, highlighting five for detailed analysis. The results show that TF-IDF remains highly effective for low-resource languages, while contextual embeddings like FastText and transformer-based document representations, such as those produced by Sentence-BERT, exhibit language-specific strengths. Principal Component Analysis (PCA) reduces training time without compromising performance, particularly benefiting FastText and neural models such as Multi-Layer Perceptrons (MLP). Computational efficiency analysis underscores the trade-off between model complexity and processing cost. Our framework provides a scalable solution for multilingual emotion detection, addressing the challenges of linguistic diversity and resource constraints.

1 Introduction

Emotion labeling in Natural Language Processing (NLP) is critical for enabling machines to better interpret human emotional expressions, fostering empathetic and context-aware AI systems. Traditional single-label emotion detection oversimplifies human affect by assigning a single dominant emotion to text, ignoring the complex spectrum of overlapping emotions often present in real-world scenarios (Plutchik, 2001). In contrast, multi-label emotion detection aligns more closely with authentic human experiences, where texts may simultaneously express multiple emotions (e.g., joy and surprise, or sadness and anger). It provides ecologically valid representations of emotional complexity, better reflecting nuanced psychological states. However, multi-label emotion detection can be challenging, such as models must account for emotion co-occurrence, resolve subtle semantic ambiguities, and avoid overfitting to sparse or imbalanced label distributions (Ekman, 1992; Wang et al., 2016; Zhang et al., 2018).

Traditional approaches to multi-label emotion detection have predominantly relied on *feature-centric frameworks* that leverage handcrafted linguistic and statistical features. While effective in monolingual settings, such frameworks often required language-specific resources (e.g., lexicons for each target language (Baccianella et al., 2010)), limiting cross-lingual scalability. These features (e.g., lexicons, syntactic patterns) often fail to model the complex interdependencies and contextual nuances required for multi-label emotion detection, as they struggle to capture dynamic label correlations and contextualized affective semantics (Baccianella et al., 2010; Mohammad et al., 2018; Bostan and Klinger, 2018).

In this paper, we conduct a comprehensive study on the development of a feature-centric framework to address the challenges of cross-lingual adaptability and multi-label emotion detection through a three-stage methodological pipeline: (1) feature extraction/document representation, (2) dimensionality reduction and (3) model training. For (1), we unify diverse feature representations ranging from interpretable shallow features (e.g., TF-IDF, Bag-of-Words) to contextually rich embeddings (e.g., FastText, BPE) and Transformer-based semantic encodings (e.g., Sentence-BERT). For (2), we reduce the dimensionality of document representations to prevent model overfitting and accelerate the subsequent step. For (3), we systematically evaluate traditional machine learning classifiers (e.g., SVM, RF) and deep learning architectures (e.g.,

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2140–2148
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

MLP) to optimize label dependency modeling. The modular design allows interchangeable classifier integration, balancing interpretability (via traditional models) and performance (via neural approaches) for diverse multilingual use cases.

The source code for this paper is publicly available on GitHub[1].

2 Background and System Overview

The development pipeline within the system comprised three distinct stages: (1) feature representation utilizing a diverse set of techniques such as TF-IDF, FastText and Sentence-Transformers, (2) dimensionality reduction of feature vectors via Principal Component Analysis (PCA), and (3) model training and prediction employing a suite of algorithms such as Decision Trees (DT) and Multi-Layer Perceptrons (MLP).

2.1 Document Representation

Raw text data in human languages are sequences of variable-length symbolic representations, which challenge machine learning algorithms that require fixed-size numeric vectors (Mikolov et al., 2013). An effective document representation is essential for optimal NLP performance. To address the complexities of multi-language data, various feature representation techniques have been explored.

2.1.1 Traditional Features

The traditional approaches represent a document d by creating a list of unique words and assigning each word w a numeric value, such as Bag of words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF).

We use scikit-learn's `CountVectorizer` to extract BoW features, which represent word occurrence in a document without considering word frequency or significance. This approach considers all words equally, including common terms like "*the*" or "*a*", which carry minimal meaningful information. In contrast, TF-IDF adjusts for the importance of words by considering both their frequency in a document (TF) and their rarity across the corpus (IDF), reducing the impact of common words. We use scikit-learn's `TfidfVectorizer`, where IDF is computed as:

$$\mathrm{IDF}(w, d) = \log \frac{1 + N}{1 + \mathrm{DF}(w)} + 1 \qquad (1)$$

where N is the total number of documents and $\mathrm{DF}(w)$ is the document frequency of w.

Preprocessing is critical in traditional feature representations, since it directly influences feature quality and model training. To adapt our model to multilingual scenarios and capture nuanced emotional expressions, we employ GemmaTokenizer (Shelpuk, 2024) in the tokenization step, instantiated with the preset "*gemma_2b_en*". Trained on a diverse multilingual corpus, GemmaTokenizer excels in handling linguistic diversity and demonstrates robust performance in multilingual tokenization tasks.

2.1.2 Pretrained Word Embeddings

Traditional methods like TF-IDF and BoW treat words as discrete, independent units, thereby completely discarding word order and local contextual information during the encoding process, leading to high-dimensional, sparse representations that limit computational efficiency and NLP task performance (Mikolov et al., 2013; Pennington et al., 2014). In addition, their heavy reliance on training corpus results in their inability to handle out-of-vocabulary (OOV) words, as they lack a mechanism to infer the representation or meaning of terms that are not present in the training corpus.

Pre-trained word embeddings offer substantial advantages in capturing semantic meaning, contextual dependencies, and generalization capabilities. Unlike traditional sparse representations, these dense vector embeddings encode rich linguistic features by leveraging large-scale corpora during pretraining, enabling them to model nuanced semantic relationships and syntactic patterns. FastText (Bojanowski et al., 2017) can capture both syntactic and semantic relationships by effectively modeling morphological structures. Byte Pair Embeddings (BPEs)(Heinzerling and Strube, 2018) decompose words into subwords, while Contextual String Embeddings (CSEs)(Akbik et al., 2018) provide context-sensitive representations, dynamically adapting to word meanings. We use the Flair NLP Toolkit[2] to extract these embeddings and `DocumentPoolEmbeddings` for aggregating word-level embeddings into document-level representations via mean pooling.

Given that the cross-lingual representation ability of most mainstream pre-trained language models remains constrained by the limited coverage of training corpora, these models often manifest

[1] https://github.com/YhzyY/SemEval2025-Task11

[2] https://github.com/flairNLP/flair

systematic representational failures when processing low-resource languages excluded from training data. This closed-corpus modeling paradigm inherently imposes significant capabilities limitations in multilingual scenarios. To address this, we leverage Large Language Models (LLMs) to assist with unseen languages by leveraging language family classification. Considering a low-resource language such as the Oromo language, it is not included in the pre-trained FastText embeddings, as outlined in the FastText documentation (Bojanowski et al., 2017). we use Baidu `Qianfan`[3] model to identify the most linguistically similar supported language in FastText. The text in unseen languages, such as Oromo, is then represented using the embeddings of the identified language. The full query is presented in Appendix A. This end-to-end self-adaptive multilingual emotion detection framework significantly enhances the system's ability to process unseen languages while fundamentally eliminating the dependencies on manually annotated language family labels and expert-curated linguistic representation rules, thus circumventing the prohibitive costs of human annotation and resolving the acute scarcity of training data and domain experts in endangered languages.

2.1.3 Transformers

Transformer-based document representations, such as those produced by Sentence-BERT (Reimers and Gurevych, 2019, 2020), leverage the Transformer architecture to selectively focus on semantically relevant segments of text, thereby enhancing feature extraction and improving representational accuracy. Our system embedded the SentenceTransformers [4] with the pretrained "*paraphrase-multilingual-mpnet-base-v2*" model, which supports over 50 languages. Document embeddings are generated using the library's encode function.

2.2 Dimensionality Reduction

The inherent high cardinality of lexical features within textual data frequently results in high-dimensional embedding, which can lead to computational challenges and model overfitting. To reduce the dimensionality of document representations, we first normalize the text to unit norm using scikit-learn's `Normalizer`. Subsequently, Principal component analysis (PCA) (Pearson, 1901) is applied to project the features into a lower-dimensional space. Both the `Normalizer` and PCA utilize default parameter settings.

2.3 Model Training

In this section, we present the methods employed for model training, encompassing traditional machine learning approaches as well as simple deep learning architectures such as MLP.

2.3.1 Traditional Machine Learning

As candidate models for training, we employ a variety of traditional machine learning algorithms, including Decision Trees (DT), k-Nearest Neighbors (KNN), Random Forest (RF) and Support Vector Machines (SVM). However, relying solely on these traditional methods often results in suboptimal accuracy (Le and Mikolov, 2014). To overcome this limitation, we adopt ensemble learning techniques, specifically constructing a majority voting classifier that aggregates predictions from a collection of base classifiers to improve overall performance.

2.3.2 Deep Learning

Multi-Layer Perceptrons (MLPs), with their multi-layered neuron architecture, can learn complex patterns in data, making them well-suited for emotion detection tasks where sentiment often depends on intricate word combinations (Goodfellow et al., 2016). To account for linguistic variations across languages, we use Grid Search to evaluate multiple parameter combinations and select the one that maximizes the *F1-macro* score, ensuring robust and accurate emotion predictions across diverse linguistic contexts.

3 Experimental Data

We use the BRIGHTER dataset (Muhammad et al., 2025a; Belay et al., 2025) provided by SemEval 2025 Task 11 Track A to conduct our experiments., It consists of human-annotated short texts in 28 languages, such as English, German and Russian. The training dataset comprises 65,098 multi-label samples, each annotated with emotion labels — anger, fear, joy, sadness, surprise, and disgust — representing the emotions most likely experienced by the speaker, as inferred from the text. We used only the provided datasets during the development and evaluation phases, no additional training data was introduced to boost the performance. Table 1 shows the statistics in selected languages, with a full list available in the Appendix B.

[3] https://qianfan.readthedocs.io/en/stable/qianfan.html

[4] https://sbert.net/

Table 1: Data Splits: Number of Train (#train), Development (#dev) and Test (#test) samples in our experiments.

Language	#train	#dev	#test
Marathi	2415	100	1000
Spanish	1996	184	1695
Hindi	2556	100	1010
Romanian	1241	123	1119
Russian	2679	199	1000

Table 2: F1-macro scores for document representations across selected languages using voting classifier.

Language	TF-IDF	FastText	SBERT
Marathi	**0.7438**	0.4277	0.6654
Spanish	**0.6561**	0.4046	0.5867
Hindi	0.4927	0.3067	**0.5682**
Romanian	0.4358	0.4486	**0.5630**
Russian	**0.7107**	0.3472	0.5767

Table 3: F1-macro scores using SBERT embeddings.

Language	DT	Voting	MLP
Marathi	0.4275	0.6654	**0.8389**
Spanish	0.5022	0.5867	**0.7076**
Hindi	0.4490	0.5682	**0.7374**
Romanian	0.5167	0.5630	**0.6375**
Russian	0.4661	0.5767	**0.7188**

4 Results

We perform experiments in 28 languages and evaluate model performance using the *F1-macro* score. Additionally, we record time consumption and generate confusion matrices to further analyze the models' performance.

4.1 Representation Selection

To investigate the impact of document representation methods on prediction outcomes, we conducted a controlled experiment with varying representations and a fixed learning algorithm.

We experimented with various representation methods, and Table 2 presents the *F1-macro* scores for the best-performing candidates from three representation approaches (i.e. traditional features, pre-trained word embeddings and transformers) across five languages[5]. To reduce the variance introduced by individual classifiers, we employ a majority voting classifier, thereby providing a more stable basis for evaluating the performance differences across various representation methods. The results of all 28 languages can be found in Appendix Section C. TF-IDF consistently outperforms other document representations across 3 out of 5 selected languages, achieving the highest F1-macro score, particularly in Marathi (0.7438). This highlights TF-IDF's effectiveness, especially in low-resource languages, as it relies on word frequency rather than pre-trained embeddings. Sentence-BERT (SBERT) shows mixed performance across languages. While it achieves the best result for Romanian (0.5630) and Hindi (0.5682), it performs worse than TF-IDF in the majority of other languages. This suggests that semantically rich contextual embeddings, such as those produced by SBERT, can offer advantages in certain linguistic contexts, but may not consistently outperform simpler lexical representations across all languages. FastText performs moderately well in some cases, such as Romanian (0.4486), but struggles in others, indicating its sensitivity to language-specific characteristics. These results emphasize that while pre-trained embeddings offer advantages in certain contexts, traditional frequency-based representations like TF-IDF remain highly competitive for emotion detection in multilingual settings.

4.2 Learning Algorithms

Using a consistent document representation, we evaluate the impact of different learning algorithms on prediction outcomes. Table 3 shows the performance of various algorithms with SBERT embeddings. The performance hierarchy is consistent across languages: MLP > Voting > DT, with MLP demonstrating the best ability to capture complex emotional patterns. The Voting classifier, combining KNN, DT and RF performs moderately, outperforming DT but lagging behind MLP. DT show weaker performance, indicating their limited capacity to model emotional nuances. These results highlight the importance of algorithm choice, with MLP being particularly effective for multilingual emotion detection.

4.3 Ablation Study

To access the contribution of individual components, we conducted an ablation study by removing specific modules and analyzing their impact on performance. The only removable component in

[5]Due to the page limit, we randomly selected five languages as examples.

Table 4: Training time in seconds with (w/) and without (w/o) PCA. The best results are bolded.

		DT	Voting	MLP
w/o PCA	TF-IDF	**1.4894**	**59.7822**	200.7861
	FastText	**0.8623**	**17.2231**	**34.3866**
	SBERT	**2.6267**	**36.9725**	**41.6211**
w/ PCA	TF-IDF	10.7575	201.4630	**114.9862**
	FastText	0.9941	23.4849	41.7187
	SBERT	2.7614	49.2546	47.1017

Table 5: F1-macro scores w/ and w/o PCA.

		DT	Voting	MLP
w/o PCA	TF-IDF	**0.5516**	**0.6561**	**0.6284**
	FastText	**0.3825**	0.4046	0.6479
	SBERT	**0.5022**	**0.5867**	0.7076
w/ PCA	TF-IDF	0.3710	0.3931	0.6100
	FastText	0.3805	**0.4407**	**0.6558**
	SBERT	0.4237	0.5038	**0.7093**

our system is the dimensionality reduction step. Using the Spanish language dataset as an example, we remove PCA to evaluate its impact on predictive performance. PCA impacts multilingual emotion detection frameworks differently based on representation-classifier pairings: for TF-IDF, it reduces MLP training time by lowering dimensionality but increases overhead for DT and the Voting classifier without accuracy gains. FastText benefits from PCA-driven noise reduction (i.e. improving accuracy) but incurs higher computational costs to retain variance, whereas SBERT's performance slightly declines as PCA strips contextual nuances critical for emotion differentiation, despite longer training times. Tree-based models such as DT remain unaffected, prioritizing raw feature hierarchies over reduced embeddings. These results emphasize that PCA's value depends on representation type (i.e. contextual vs. static) and classifier architecture, advocating for selective use to optimize multilingual systems.

4.4 Data Imbalance

Data imbalance in the training dataset can significantly impact model performance, causing bias towards the majority class and resulting in poor predictions for the minority class.

For the Hindi dataset trained with FastText and MLP, over 78% of 2,556 samples are labeled as not "*anger*" leading to a high specificity score (0.9405) but a low recall score (0.5625) for the "*anger*" label,

Table 6: Confusion matrix on Hindi language subset.

	anger	disgust	fear	joy	sadness	surprise
TP	0.5625	0.5000	0.6429	0.6364	0.2941	0.6667
TN	0.9405	1.0000	0.9651	0.9438	0.9277	0.9780
FP	0.0595	0.0000	0.0348	0.0562	0.0723	0.0220
FN	0.4375	0.5000	0.3571	0.3636	0.7059	0.3333

as shown in Table 6. A similar pattern is observed for other emotion labels, highlighting that imbalance between positive and negative samples can undermine prediction accuracy. These results emphasize the significant impact of label distribution on system performance.

4.5 Model Efficiency

The choice of document representation and learning algorithm significantly affects the computational efficiency of emotion detection systems, as demonstrated by comparing the efficiency of time over five languages using FastText embeddings in Table 7. Simpler models, such as DT, exhibit rapid training speeds (0.50–1.14s), making them computationally efficient but often at the expense of accuracy. In contrast, MLP achieves superior predictive performance but requires significantly longer training times (24.84–53.01s), representing a substantial increase in computational cost over DT. The Voting classifier, which integrates multiple models, falls between these extremes, with training times ranging from 9.67s to 23.51s. Despite these differences in training efficiency, all models achieve sub-millisecond inference speeds, with prediction times between 0.3 ms and 0.7 ms, except for Voting in Marathi (0.95 µs) and Russian (1.92 µs). This suggests that inference latency is primarily influenced by model architecture rather than language complexity. These results highlight key efficiency-accuracy trade-offs. While high-dimensional embeddings like TF-IDF achieve strong F1 scores (e.g., Marathi: TF-IDF 0.68), their computational costs (140.24s) may be prohibitive for real-time applications. FastText with MLP provides a balanced alternative, offering competitive accuracy (Marathi: 0.67) with moderate computational cost (embedding: 1.00s, training: 43.56s), underscoring the need for multi-objective optimization in multilingual emotion detection.

5 Conclusions

This study presents a feature-centric framework for cross-lingual multi-emotion detection in short

Table 7: Train and test time in seconds using FastText embeddings.

Language	DT		Voting		MLP	
	Train	Test	Train	Test	Train	Test
Marathi	1.06	6e-4	23.51	9.54e-7	43.56	6e-4
Spanish	0.86	4e-4	17.22	0.0000	34.39	7e-4
Hindi	1.14	3e-4	22.26	0.0000	49.78	7e-4
Romanian	0.50	3e-4	9.67	0.0000	24.84	6e-4
Russian	1.14	4e-4	22.69	1.92e-6	53.01	7e-4

texts, designed to dynamically adapt of document representations and learning algorithms for optimal language-specific performance. Through a comprehensive comparative study across 28 languages—highlighting five for demonstration—we evaluate three key components: document representation, dimensionality reduction, and model training. Our findings show that the proposed pipeline is adaptable across languages with minimal adjustments, effectively balancing computational efficiency and detection accuracy. Experimental results validate its robustness in multi-label emotion prediction, particularly for low-resource languages.

In future work, we plan to refine the framework by optimizing feature-classifier selection for each language, leveraging advanced LLMs for enhanced feature extraction, and training FastText word vectors to improve representation quality, particularly for low-resource languages. We will also focus on determining optimal PCA configurations and evaluating its impact on the performance across different languages and emotion categories to ensure the robustness and reliability of our framework.

Ethical Statements

This paper presents a feature-centric framework for cross-lingual multi-emotion detection, utilizing the publicly available BRIGHTER dataset (Muhammad et al., 2025a). While leveraging its multilingual resources, we explicitly acknowledge that emotional expression is culturally and linguistically dependent, which may introduce biases in annotation and model predictions, particularly in capturing nuanced emotional expressions across low-resource languages and dialects. To address these challenges, we advocate for responsible development and deployment of the framework, emphasizing ongoing research into bias detection, fairness in cross-lingual emotion analysis, and mitigation of potential data-driven biases. These efforts aim to ensure equitable and ethical applications of the technology while transparently addressing its cultural and linguistic limitations.

Acknowledgments

We would like to thank the organizers of SemEval 2025 Task 11 for hosting the shared task and providing valuable resources that enabled this research. Their efforts in promoting multilingual emotion analysis are greatly appreciated. We are also grateful to the anonymous reviewers for their insightful feedback and constructive suggestions, which have helped us improve the quality and clarity of this work.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 2200–2204.

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.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Laura A. Bostan and Roman Klinger. 2018. An Analysis of Annotated Corpora for Emotion Classification in Text. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 2104–2119. Association for Computational Linguistics.

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

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

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1188–1196.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval)*, pages 1–17. Association for Computational Linguistics.

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 Zhang, 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, 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. 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.

Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

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. 2020. 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.

Shelpuk. 2024. Llm practitioner's guide: Gemma, a game-changing multilingual llm.

Wenbo Wang, Lei Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2016. Emotional attention detection with multi-label classification. *IEEE Transactions on Affective Computing*, 9(4):468–480.

Linhao Zhang, Shuai Wang, and Bing Liu. 2018. Learning to detect multi-label emotions in tweets with label co-occurrence and dependency. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3299–3309.

A LLM query for Language Family Classification

The LLM employed in our system is Baidu `Qianfan` "ernie-4.0-8k-latest", and the prompt is `"You are a linguist working on language classification and are familiar with the given languages: (list the known languages). Please select the language from the list that is most similar to (given language) based on language family and geographic distance in terms of population distribution."`.

B Data Statics

Table 8 presents the number of training, development and test splits across all 28 languages in our experiments, sorted by number of training samples in descending order.

Table 8: Train (train), Development (dev) and Test (test) splits for all 28 languages.

Language	#train	#dev	#test
Nigerian pidgin	3728	620	1870
Tigrinya	3681	614	1840
Amharic	3549	592	1774
Oromo	3442	574	1721
Somali	3392	566	1696
Swahili	3307	551	1656
Yoruba	2992	497	1500
Igbo	2880	479	1444
English	2768	116	2767
Russian	2679	199	1000
Chinese	2642	200	2642
German	2603	200	2604
Hindi	2556	100	1010
Ukrainian	2466	249	2234
Kinyarwanda	2451	407	1231
Marathi	2415	100	1000
Portuguese(Brazilian)	2226	200	2226
Hausa	2145	356	1080
Spanish	1996	184	1695
Moroccan Arabic	1608	267	812
Makhuwa	1551	258	777
Portuguese(Mozambican)	1546	257	776
Romanian	1241	123	1119
Afrikaans	1222	98	1065
Swedish	1187	200	1188
Tatar	1000	200	1000
Sundanese	924	199	926
Algerian Arabic	901	100	902

C Model Performance

Table 9 shows the F1-macro scores across all 28 languages using three representation methods (TF-IDF, FastText and SBERT) combined with three classifiers (DT, Voting and MLP). The languages are listed in descending order based on the number of training samples, aligning with the organizational schema of Appendix B.

Overall, the transformer-based document representations, such as Sentence-BERT, generally outperform TF-IDF and FastText across 18 out of 28 languages, demonstrating their effectiveness in capturing semantic nuances. In contrast, the pretrained word embeddings such as FastText tend to yields lower scores in most cases, likely due to the limited representation of several low-resource or less commonly used languages in its pretraining corpus, resulting in suboptimal embedding quality for these languages. On the classifier side, the deep learning model MLP consistently delivers the best performance, particularly when combined with SBERT representation, highlighting the advantage of deep learning models in leveraging dense contextual representations. Traditional machine learning approaches, such as Decision Tree, performs better with sparse features like TF-IDF but struggles with dense representations. These findings collectively underscore the importance of combining semantically rich representations with expressive classifiers for robust multilingual emotion detection.

Table 9: F1-macro scores for all 28 languages.

Language		DT	Voting	MLP
Nigerian pidgin	TF-IDF	0.3268	0.2764	0.3457
	FastText	0.2681	0.2600	0.3990
	SBERT	0.3065	0.2721	0.4142
Tigrinya	TF-IDF	0.2473	0.1758	0.2846
	FastText	0.0477	0.0430	0.0208
	SBERT	0.1957	0.1791	0.1969
Amharic	TF-IDF	0.2281	0.2557	0.3418
	FastText	0.0569	0.1216	0.0142
	SBERT	0.3014	0.3111	0.4121
Oromo	TF-IDF	0.3222	0.3351	0.4172
	FastText	0.1756	0.2046	0.1033
	SBERT	0.1974	0.1969	0.2397
Somali	TF-IDF	0.2481	0.2704	0.3523
	FastText	0.1372	0.0756	0.1296
	SBERT	0.1460	0.1408	0.2577
Swahili	TF-IDF	0.2015	0.1352	0.1966
	FastText	0.1563	0.0653	0.1710
	SBERT	0.1469	0.0913	0.1442
Yoruba	TF-IDF	0.2056	0.2085	0.2610
	FastText	0.1495	0.1279	0.1899
	SBERT	0.1432	0.0782	0.1989
Igbo	TF-IDF	0.3789	0.4140	0.4888
	FastText	0.2321	0.2867	0.3753
	SBERT	0.2591	0.2809	0.3671
English	TF-IDF	0.3302	0.3908	0.5197
	FastText	0.4177	0.4069	0.6139
	SBERT	0.4421	0.5683	0.6654
Russian	TF-IDF	0.6418	0.7107	0.7155
	FastText	0.3202	0.3472	0.6341
	SBERT	0.4661	0.5767	0.7188
Chinese	TF-IDF	0.2730	0.2964	0.3889
	FastText	0.0904	0.1207	0.0711
	SBERT	0.3475	0.3831	0.5402
German	TF-IDF	0.3054	0.2889	0.3611
	FastText	0.3116	0.3101	0.4064
	SBERT	0.3130	0.3535	0.5011
Hindi	TF-IDF	0.5092	0.4927	0.6122
	FastText	0.3261	0.3067	0.6051
	SBERT	0.4490	0.5682	0.7374
Ukrainian	TF-IDF	0.2751	0.2906	0.2735
	FastText	0.1384	0.1123	0.2092
	SBERT	0.2675	0.2874	0.4472

Language		DT	Voting	MLP
Kinyarwanda	TF-IDF	0.2779	0.2595	0.3163
	FastText	0.2272	0.1815	0.1751
	SBERT	0.1834	0.2146	0.3432
Marathi	TF-IDF	0.6817	0.7438	0.7640
	FastText	0.3849	0.4277	0.6711
	SBERT	0.4275	0.6654	0.8389
Portuguese (Brazilian)	TF-IDF	0.2170	0.1723	0.2296
	FastText	0.2004	0.1648	0.3198
	SBERT	0.2821	0.2681	0.5131
Hausa	TF-IDF	0.4355	0.5250	0.5560
	FastText	0.3225	0.3250	0.3538
	SBERT	0.3065	0.3445	0.4667
Spanish	TF-IDF	0.5516	0.6561	0.6284
	FastText	0.3825	0.4046	0.6479
	SBERT	0.5022	0.5867	0.7076
Moroccan Arabic	TF-IDF	0.2142	0.2014	0.2838
	FastText	0.1942	0.2182	0.3748
	SBERT	0.2530	0.3409	0.4183
Makhuwa	TF-IDF	0.2055	0.1125	0.1554
	FastText	0.0869	0.0393	0.0640
	SBERT	0.1366	0.0556	0.0985
Portuguese (Mozambican)	TF-IDF	0.2651	0.1977	0.1571
	FastText	0.1514	0.1009	0.3009
	SBERT	0.1794	0.2147	0.4021
Romanian	TF-IDF	0.4484	0.4358	0.6110
	FastText	0.4633	0.4486	0.6217
	SBERT	0.5167	0.5630	0.6375
Afrikaans	TF-IDF	0.2153	0.2366	0.1813
	FastText	0.1763	0.1155	0.1283
	SBERT	0.2655	0.2834	0.4729
Swedish	TF-IDF	0.3415	0.2694	0.2542
	FastText	0.2619	0.2715	0.3729
	SBERT	0.3144	0.3118	0.4310
Tatar	TF-IDF	0.4457	0.4724	0.4528
	FastText	0.2508	0.2339	0.3885
	SBERT	0.2418	0.2385	0.3777
Sundanese	TF-IDF	0.3503	0.3642	0.3690
	FastText	0.2455	0.2104	0.2452
	SBERT	0.2407	0.2851	0.3605
Algerian Arabic	TF-IDF	0.3464	0.3311	0.4770
	FastText	0.3109	0.3343	0.4421
	SBERT	0.3642	0.3782	0.5126