INFOTEC-NLP at SemEval-2025 Task 11: A Case Study on Transformer-Based Models and Bag of Words

Emmanuel Santos[†] and **Mario Graff**[†,‡]
[†] INFOTEC Centro de Investigación e Innovación en
Tecnologías de la Información y Comunicación, Aguascalientes, México
[‡] Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), México
e.ss.rdz@gmail.com
mario.graff@infotec.edu.mx

Abstract

Emotion detection in text is a key task in computational linguistics, challenged by linguistic ambiguities, cultural differences, and the scarcity of non-English resources, limiting its multilingual applicability. While pre-trained transformers and neural networks have shown strong performance, research remains largely English-centric, highlighting the need for inclusive, cross-linguistic approaches. This work tackles SemEval-2025 Task 11, Track A: Multi-label Emotion Detection (Muhammad et al., 2025b), predicting perceived emotions (joy, sadness, fear, anger, surprise or disgust) in text snippets. We propose a hybrid model combining XLM-RoBERTa embeddings with Bi-LSTM and multi-head attention, enhancing contextual understanding and classification across languages. Experiments on the task dataset show our model effectively captures emotional nuances, outperforming the baselines in most languages. Results show that our method improves macro-F1 scores for multilingual emotion classification. These findings highlight the value of combining transformer-based embeddings with structured sequence modeling to better represent linguistic and cultural diversity.

1 Introduction

Since its inception, artificial intelligence has sought to solve human and social problems through techniques such as natural language processing (NLP), which combines computational and linguistic methods to enable computers to understand human languages in formats such as text and audio/voice (Acheampong et al., 2020).

Several initiatives have introduced emotion detection tasks to encourage the research community to develop competitive tasks for processing, understanding, and generating text. These efforts present new research challenges and establish state-of-the-art results in this field. There are works that have significantly advanced the field of emotion detection by tackling different aspects, e.g., Mohammad et al. (2018) laid a strong foundation by introducing multi-label emotion classification and emotion intensity prediction in tweets, providing a benchmark for fine-grained affect detection. Building on this, the work of Kumar et al. (2024) extends the challenge to conversational contexts, emphasizing emotion shifts and their reasoning, a crucial step toward developing systems capable of understanding emotional transitions. Meanwhile, García-Vega et al. (2020) contribute to the field by introducing emotion detection in Spanish-language tweets, addressing the gap in multilingual emotion classification, and enhancing NLP applications for non-English texts. These joint efforts highlight the growing importance of emotion-aware systems, particularly in social media monitoring, mental health applications, and human-computer interaction (Al-Saqqa et al., 2018). By broadening the scope of affective computing to include emotion intensity, conversational reasoning, and language diversity, these advancements improve the depth and applicability of the field. Their contributions drive the development of more context-aware, explainable, and adaptable models.

Despite significant advancements in recent years to address the challenges of sentiment analysis, emotion classification remains a complex task for NLP systems, whose relevance has continued to grow over time.

This paper presents a model for Track A of SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection (Muhammad et al., 2025a; Belay et al., 2025; Muhammad et al., 2025b), which focuses on predicting a speaker's perceived emotions based on a given text snippet. We propose to solve the challenge as a multi-label classification problem using an approach that leverages a transformer-based encoder to extract deep semantic representations, with a Bi-LSTM for cap-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 350–356

turing temporal dependencies and multi-head attention mechanisms to refine emotion-specific features further. This architecture enables the effective detection of multiple emotions in multilingual text, contributing to advancements in affective computing.

The rest of the paper is organized as follows: Section 2 describes the problem and the related work. Section 3 presents the system proposed and discusses the experimental setup to tackle task A; meanwhile, Section 4 analyzes the results. The conclusions are presented in Section 5.

2 Background

Emotion detection involves identifying a person's sentiments or feelings. In computational linguistics, it identifies a discrete emotion in a text (Nandwani and Verma, 2021). This task is challenging due to cultural differences, linguistic ambiguity, and the use of slang.

The techniques for emotion detection include lexicon-based, machine learning, and deep learning approaches (Seyeditabari et al., 2018). Lexicon-based approaches use dictionaries of words with sentiment values to determine the predominant emotion in a text. On the other hand, machine learning models rely on labeled datasets to train supervised models that predict emotions. The most recent methodologies are based on deep learning, which applies neural networks with minimal feature engineering.

The work of Ameer et al. (2023) tackles the problem of multi-label emotion classification using an approach that combines multiple attention mechanisms with recurrent neural networks and pre-trained transformer models (through transfer learning). This approach achieves results that surpass the state of the art in terms of accuracy. Wang et al. (2024) propose an emotion detection model that distills knowledge from a high-performing English monolingual model to a multilingual model. This approach enhances emotion detection, demonstrating the effectiveness of knowledge transfer for robust and explainable multilingual models.

Despite these advancements, emotion recognition research has primarily focused on high-resource languages, leaving low-resource languages underrepresented. Nevertheless, recently, two studies have introduced large-scale multi-label emotion datasets aimed at improving multilingual emotion classification. Muhammad et al. (2025a) introduce BRIGHTER, a collection of emotion-annotated datasets spanning 28 languages from Africa, Asia, Eastern Europe, and Latin America. The dataset integrates diverse sources and employs various annotation strategies by fluent speakers. It features multi-label annotations and intensity levels, enabling a more nuanced understanding of emotions. Experimental results highlight the variability in large language model (LLM) performance across languages and explore cross-lingual transfer learning. Findings show that multilingual models achieve better results when trained on linguistically related languages, while LLMs performance drops significantly in low-resource settings.

Belay et al. (2025) introduce EthioEmo, a multi-label emotion dataset for Ethiopian languages, compiled from diverse sources such as social media and news. The study extensively evaluates multiple language model architectures and explores translation-based evaluation methods. The findings highlight the challenges of multi-label emotion classification in low-resource settings and expose the limitations of LLMs in capturing emotional nuances across linguistic structures.

These datasets and experiments advance the state-of-the-art in multilingual and multi-label emotion classification as part of SemEval 2025 Task 11.

3 System Overview

The proposed system is designed for multilingual emotion classification using a hybrid deep learning approach. It integrates a pre-trained transformer-based model (Vaswani et al., 2017), a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) (Bi-LSTM) network, and a multi-head attention mechanism (Vaswani et al., 2017).

The decision to integrate Bi-LSTM and attention mechanisms with transformer-based architectures stems from the complementary advantages each component brings to natural language processing tasks. Bi-LSTM is well suited for capturing sequential dependencies and maintaining word order, both of which are crucial in tasks such as text classification and sentiment analysis, where the temporal structure of input matters. Pure transformer models can struggle with fine-grained temporal relationships (particularly in low-resource settings), as they lack inherent sequence modeling capabilities (Otter et al., 2020). Incorporating Bi-LSTM allows the architecture to preserve these structural cues, while

the attention mechanism adds the ability to dynamically highlight important words or phrases, boosting both accuracy and interpretability. This synergy has been demonstrated in several contexts. For instance, combining Bi-LSTM with BERT has shown improved entity recognition in medical text due to better handling of position-sensitive features (Zalte and Shah, 2024), and sentiment analysis in social media content benefits from this hybrid by capturing emotional nuances more effectively (Bader et al., 2024). Attention-enhanced Bi-LSTM architectures also contribute significantly to explainability, an increasingly vital consideration in modern NLP systems (Galassi et al., 2020). Additionally, in scenarios such as microblog sentiment classification, this hybrid approach helps address overgeneralization issues commonly seen in transformer-only models (Jia, 2022). Altogether, the combination of Bi-LSTM, attention, and transformer components provides a well-rounded solution that balances semantic depth, sequential awareness, and interpretability, making it a strong alternative to standalone transformer models.

Below is an overview of the system components and their functionalities:

- Transformer-based model as a feature extractor: The system uses a pre-trained transformer-based model (e.g., BERT, XLM-RoBERTa, etc.) to generate contextualized embeddings for input text sequences.
- Bidirectional LSTM: The hidden states from the transformer are passed through bidirectional LSTM to capture sequential dependencies in the text.
- Multihead attention mechanism: The bi-LSTM outputs are passed through a multi-head attention layer to focus on the most relevant parts of the sequence.
- Classification head: The attention outputs are averaged across the sequence length to obtain a fixed-size contextual representation. This representation is then passed through a fully connected layer to produce logits for each emotion class. A sigmoid activation is applied to the logits to obtain probabilities for multi-label classification.

Finally, predictions are binarized using a threshold, where values greater than the threshold are classified as positive labels.

To evaluate the effectiveness of our proposed system, we trained multiple variants using different transformer-based models as feature extractors. Our experimental setup aims to assess how the choice of transformer architecture influences multilingual emotion classification performance, specifically focusing on their impact on learned representations and downstream classification capabilities. In the following paragraphs, we describe our experiments' specific model configurations and dataset.

We trained two model variants. The first leverages XLM-RoBERTa (XLM-R) (Conneau et al., 2019) as the feature extractor for all languages in the dataset, while the second employs AfriBERTa (Ogueji et al., 2021), specifically for African languages. For each, we compared the performance of the base and large configurations. We evaluate our models on a multilingual emotion classification dataset annotated for six emotion categories. Each instance can have multiple emotion labels, making this a multi-label classification task.

Datasets The models are trained on SemEval-2025 Task 11 (track A) datasets which contain text samples labeled with six emotions: joy, sadness, fear, anger, surprise, and disgust for 29 languages (Muhammad et al., 2025a; Belay et al., 2025). Each text snippet is annotated with binary labels indicating the presence or absence of each emotion. For English and Afrikaans, the dataset includes only five emotions. This variation is handled during preprocessing to ensure compatibility with the models. The dataset is divided into training, validation and test sets.

Preprocessing Input text sequences are tokenized using the tokenizer associated with the pre-trained transformer model. Sequences are truncated or padded to a fixed length of 128 tokens. On the other hand, for languages with fewer emotions, the missing emotion is set to 0 for all samples.

Model configuration

- Pre-trained transformer encoder: Pre-trained transformer models, namely XLM-RoBERTa-base, XLM-RoBERTa-large, AfriBERTa-base and AfriBERTA-large, are used as feature extractors. The transformer generates contextualized embeddings with the size specified in its

configuration. This size is then used as the input dimension for a bidirectional LSTM

- BiLSTM layer: A single-layer bidirectional LSTM with a hidden size of 256.
- Multi-head attention: Uses 8 attention heads, with an embedding dimension of 512.
- Classification head: A fully connected layer maps the final contextual representation to six emotion logits, followed by a sigmoid activation for multi-label classification.

Training The training set is used to optimize model parameters via Binary Cross-Entropy (BCE) loss, employing the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 2×10^{-6} over 10 epochs. A batch size of 8 is used consistently for both training and evaluation across all models. Model selection is guided by the minimum training loss: the model is checkpointed whenever a lower loss is achieved in subsequent epochs. The validation set serves to monitor performance during training and supports the selection of the optimal model configuration.

Evaluation Predictions are converted to binary labels using a threshold of 0.5. For example, a predicted probability greater than 0.5 is considered the presence of the emotion (1), and lower or equal to 0.5 is regarded as the absence (0). The performance of the models is evaluated on the test set using the F1 score, computed by comparing the predicted labels against the gold-standard annotations. We include a Bag-of-Words (BoW) baseline for comparison to provide a more comprehensive assessment. The BoW follows the approach described in (Tellez et al., 2017)[1]; we set all the characters to lowercase, removing diacritics and punctuation symbols. Additionally, the users and the URLs were removed from the text. We use several tokenizers, i.e., bigrams, words, and q-grams of characters with $q = \{2, 3, 4\}$.

4 Results

The systems' performance analysis starts with the information presented in Table 1. The table re-

[1] evomsa.readthedocs.io/en/docs/bow.html

Language	base	large	BoW
afr	0.4120	**0.5094**	0.2695
amh	0.6529	**0.6939**	0.5697
arq	0.5084*	**0.5170**	0.4699
ary	0.4579	**0.5551**	0.4063
chn	0.5817	**0.6480**	0.4658
deu	0.5945	**0.6635**	0.4539
eng	0.6586	**0.7091**	0.5049
esp	0.7471	**0.7925**	0.6905
hau	0.6104	**0.6448**	0.6320*
hin	0.8616	**0.8942**	0.7588
ibo	0.5005	0.4995	**0.5555**
kin	0.3174	0.3533	**0.4468**
mar	0.8293	**0.8800**	0.7583
orm	0.4990	0.5515*	**0.5719**
pcm	0.5366	**0.5793**	0.4707
ptbr	0.5050	**0.5585**	0.3456
ptmz	0.4463*	**0.4847**	0.2516
ron	0.6877	**0.7398**	0.6146
rus	0.8272	**0.8717**	0.7224
som	0.3919	**0.4888**	0.4689*
sun	0.3977	**0.4291**	0.3803
swa	0.2564*	**0.2808**	0.2389
swe	0.5403	**0.5942**	0.4147
tat	0.6633*	**0.6809**	0.5746
tir	0.4636	0.4930*	**0.5101**
ukr	0.5640	**0.6601**	0.3756
vmw	0.1226	0.0405	**0.2626**
yor	0.1891	0.1529	**0.3716**

Table 1: Macro F1 scores were obtained in the test set per language for XLM-RoBERTa-based (base and large) models and Bag of Words (BoW). The highest performance scores are highlighted in bold. The asterisk indicates that the difference in performance is not statistically significant (5%) with respect to the best performance.

Language	Afriberta-base	Afriberta-large
amh	0.5767	**0.6123**
hau	0.6510	**0.6554**
ibo	0.4728	**0.4747**
kin	**0.4383**	0.4281
orm	0.5391	**0.5414**
pcm	0.4594	**0.4677**
som	0.4037	**0.4319**
swa	0.1528	**0.1770**
tir	**0.4656**	0.4629
yor	0.2682	**0.2757**

Table 2: Macro F1 scores obtained in test set per language for AfriBERTa-based models. The highest performance scores are highlighted in bold.

ports macro-F1 scores per language for the XLM-RoBERTa-based classifiers and the BoW baseline. The table highlights in boldface the best scores and indicates with an asterisk those scores whose difference to the best is not statistically significant (5%).[2]

Based on the results, we observe that the large model achieves a higher macro F1 score across most languages than the base model and baseline, indicating a consistent improvement in classification performance. Languages such as mar, rus, esp, eng, hin, and ron achieve high scores for both base and large models, suggesting that they benefit from better representation and greater resource availability in the training data. Several languages show moderate improvements when scaling from base to large but still lag behind the best-performing languages. It is also worth noting that languages such as ibo, yor, and vmw do not benefit from scaling, as their scores tend to decrease.

In some languages (yor, vmw, tir, som, orm, kin, ibo, and hau), the BoW outperforms or is statistically equivalent to the large model, which may indicate insufficient data to fine-tune the MLM.

Table 2 shows the F1 scores per language for the AfriBERTa-based models. According to the data, in most cases, the AfriBERTa-large model achieves slightly higher macro F1 scores than the AfriBERTa-base model. This is especially notable in languages such as amh, som and saw. However, some languages show little to no difference between the base and large models. This is the case for gbo, tir, hau and orm. On the other hand, languages such as swa and yor show the lowest overall scores; this might indicate there is a great challenge in handling these languages.

When comparing the predictions from XLM-RoBERTa-based models with those from AfriBERTa-based models, we observe that XLM-RoBERTa generally outperforms AfriBERTa in most cases. This is particularly evident for amh, hau, ibo, orm, pcm, and som, where XLM-R-large scores higher than AfriBERTa-large. However, there are exceptions where AfriBERTa slightly surpasses XLM-RoBERTa. This is the case for hau, where AfriBERTa-large marginally outperforms XLM-R-large. For yor and kin, the base and large versions of AfriBERTa outperform their XLM-R counterparts. It is worth noting, however, that swa performs worse across all AfriBERTa versions. Additionally, AfriBERTa-based models fail to outperform the baseline scores, except for amh and hau. This suggests that our AfriBERTa-based approach may not be well-suited for African languages, as it struggles to outperform the baseline consistently. As the results obtained by the XLM-R-large-based model were the highest performing, these model's predictions were submitted to the competition.

[2]The p-values of the difference in performance were estimated using bootstrap with the library described in Nava-Muñoz et al. (2024).

5 Conclusions

This study evaluates the performance of transformer-based models for multilingual emotion classification, comparing XLM-RoBERTa and AfriBERTa across a diverse set of languages. Several key takeaways emerge from our analysis.

First, model scaling improves classification performance, with large variants (XLM-R-large and AfriBERTa-large) consistently outperforming their base counterparts. XLM-RoBERTa achieves higher F1 scores than AfriBERTa in most cases, particularly for amh, hau, ibo, orm, pcm, and som, indicating that XLM-RoBERTa's multilingual pretraining provides more robust representations. However, AfriBERTa outperforms XLM-RoBERTa for specific languages such as yor and kin, suggesting that AfriBERTa's training data may better capture linguistic characteristics of certain African languages. Interestingly, scaling does not always lead to improved performance, as some languages (ibo, yor, and vmw) experience a decrease in scores when moving from base to large models. This suggests that simply increasing model capacity does not necessarily enhance classification performance for all languages, possibly due to data sparsity or overfitting. Among all evaluated models, XLM-R-large emerges as the best-performing approach for multilingual multi-label emotion detection, making it a strong candidate for robust and scalable NLP applications.

Despite its design for African languages, AfriBERTa fails to outperform the baseline in multiple cases, with the exception of amh and hau. This raises concerns about its adequacy as feature extractor for our system, specially for underrepresented languages. Furthermore, high-resource languages such as eng, esp, rus, hin, mar, and ron achieve significantly higher scores, benefiting from well-established pretraining data, whereas low-resource languages such as vmw, yor, swa, kin, som and sun

exhibit weaker performance, likely due to limited resources.

We also emphasize the importance of initiatives such as SemEval in advancing emotion detection, affective computing, and broader NLP challenges. Tasks like SemEval-2025 Task 11 provide high-quality annotated datasets that enable the systematic development and evaluation of models across diverse languages and domains. By fostering both competition and collaboration, these shared tasks drive the advancement of more robust models and methodologies. Additionally, they help uncover limitations in existing approaches, particularly for low-resource languages, and encourage research efforts toward more inclusive and generalizable NLP systems.

Our findings highlight the persistent challenges in modeling African languages for emotion classification. Future work should explore domain-adaptive pretraining, data augmentation techniques, and specialized architectures to improve multilingual and low-resource language performance

References

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

Samar Al-Saqqa, Heba Abdel-Nabi, and Arafat Awajan. 2018. A survey of textual emotion detection. In *2018 8th International Conference on Computer Science and Information Technology (CSIT)*, pages 136–142. IEEE.

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.

Mohamed Bader, Ismail Shahin, and Abdelfatah Ahmed. 2024. Covid-19 tweets analysis using hybrid bi-lstm and transformer models. In *2024 Advances in Science and Engineering Technology International Conferences (ASET)*, pages 1–6. IEEE.

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.

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.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2020. Attention in natural language processing. *IEEE transactions on neural networks and learning systems*, 32(10):4291–4308.

Manuel García-Vega, Manuel Carlos Diáz-Galiano, MA García-Cumbreras, Flor Miriam Plaza Del Arco, Arturo Montejo-Raéz, Salud María Jiménez-Zafra, E Martínez Cámara, César Antonio Aguilar, Marco Antonio Sobrevilla Cabezudo, Luis Chiruzzo, et al. 2020. Overview of tass 2020: Introducing emotion detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) Co-Located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain*, pages 163–170.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Keliang Jia. 2022. Sentiment classification of microblog: A framework based on bert and cnn with attention mechanism. *Computers and Electrical Engineering*, 101:108032.

Shivani Kumar, Md Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. 2024. Semeval 2024–task 10: Emotion discovery and reasoning its flip in conversation (ediref). *arXiv preprint arXiv:2402.18944*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

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.

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

Sergio Nava-Muñoz, Mario Graff, and Hugo Jair Escalante. 2024. Analysis of systems' performance in natural language processing competitions. *Pattern Recognition Letters*.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel W Otter, Julian R Medina, and Jugal K Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624.

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. 2018. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*.

Eric S. Tellez, Sabino Miranda-Jiménez, Mario Graff, Daniela Moctezuma, Ranyart R. Suárez, and Oscar S. Siordia. 2017. A simple approach to multilingual polarity classification in Twitter. *Pattern Recognition Letters*, 94:68–74.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.

Jaya Zalte and Harshal Shah. 2024. Contextual classification of clinical records with bidirectional long short-term memory (bi-lstm) and bidirectional encoder representations from transformers (bert) model. *Computational Intelligence*, 40(4):e12692.