IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 October 2025, accepted 11 October 2025, date of publication 20 October 2025, date of current version 4 November 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3623629

RESEARCH ARTICLE

ApuEmo: Emotion Classification in Spanish Through a Hybrid Model With Transformer and Recurrent Layer

YONATAN MAMANI-COAQUIRA[1], MANUEL J. IBARRA-CABRERA[1], MARIO AQUINO-CRUZ[1], AND WILSON J. MOLLOCONDO-FLORES[2]

[1]Informatics and Systems Department, Universidad Nacional Micaela Bastidas de Apurimac, Tamburco, Abancay 03001, Peru
[2]Basic Sciences Department, Universidad Nacional Micaela Bastidas de Apurimac, Tamburco, Abancay 03001, Peru

Corresponding author: Yonatan Mamani-Coaquira (ymamani@unamba.edu.pe)

This work was supported by the National University Micaela Bastidas of Apurimac.

ABSTRACT Emotion classification in social networks is a crucial task, driven by the increasing need to analyze the opinions and feelings expressed across various platforms such as Facebook, YouTube, Instagram, and X. This work presents a novel hybrid approach for emotion classification in Spanish-language texts, integrating the pre-trained SaBERT embedding with recurrent neural networks and attention mechanisms. A rigorous evaluation using the TASS 2020 dataset from the Workshop on Semantic Analysis for Task 2: Emotion Detection, alongside a collection of Spanish comments sourced from Facebook related to the Apurimac region in Peru, was conducted. The results show that the proposed model outperforms representative state-of-the-art models, such as ELiRF-UPV and UMUTeam, achieving a maximum F1-Macro value of 0.49. Moreover, complementary lexical and emotional analyses allowed for validating the model's behaviour in regional contexts, revealing an emotional distribution consistent with the cultural and linguistic content of the Apurimac region in Peru.

INDEX TERMS Emotion classification, Spanish language, transformers, recurrent neural networks.

I. INTRODUCTION

In recent years there is research on sentiment and emotion analysis in text has significantly increased due to the vast amounts of text available from the internet and social networks [1]. The vast amount of data generated daily has made textual content a crucial source for understanding emotions, opinions, and social trends. From social media posts to forum comments and product reviews, the emotions expressed online offer valuable insights for various fields, including psychology, politics, marketing, economics, commerce, education, and health [2]. However, the rapid growth of information also presents challenges, such as the spread of misinformation and the difficulty in distinguishing accurate content from that which could negatively affect public perception. Given the influence that online emotions can have on decision-making and the reputation of brands and institutions.

Textual comments on social networks provide a platform for users to freely express a range of emotions, including joy, surprise, anger, disgust, sadness, and fear, among others [3], [4]. Identifying and accurately classifying these emotions in comments is essential for understanding social dynamics and collective moods, which can significantly influence strategic decision-making. For instance, in marketing, analyzing positive or negative emotions related to products and services allows companies to adjust their advertising campaigns and enhance consumer perception. Similarly, in the political sphere, assessing collective emotions can yield insights into electoral trends or public reactions to specific policies. Additionally, in public health, understanding the emotions conveyed on social networks can aid in the early detection of collective emotional crises, enabling rapid and effective responses to critical situations.

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

Emotion classification in social networks is a crucial task, driven by the increasing need to analyze the opinions and feelings expressed across various platforms such as Facebook, YouTube, Instagram, and X. A review of the literature reveals that a significant portion of existing studies has primarily focused on textual comments in English for emotion and sentiment classification. This focus limits the scope and applicability of the findings to other languages, such as Spanish [2]. A detailed examination of the scientific literature indicates that the development of models specifically tailored to the Spanish language remains scarce [2], [5], [6]. This limitation is largely due to the lack of Spanish linguistic resources, including corpora and specialized pretrained models, which has been identified as a primary challenge in recent studies [7], [8]. On the other hand, deep learning techniques have greatly advanced this area of research, particularly through architectures that incorporate recurrent connections, convolutional layers, and attention mechanisms [9], [10], [11]. Additionally, pre-trained models for the Spanish language such as BETO [12], SaBERT [13], RoBERTuito [14], and BERTin [15] and MarIA [16], which are based on Bidirectional Encoder Representations from Transformers (BERT) [17], have shown competitive performance in sentiment and emotion classification tasks. These approaches have shown significant potential for modeling temporal dependencies and extracting spatial patterns present in the texts. Furthermore, hybrid models that integrate linguistic elements with deep neural networks have notably enhanced the ability to identify and classify complex emotional expressions [18], [19], [20], [21].

This work presents a novel hybrid model, named ApuEmo, specifically designed for emotion classification in Spanish-language texts, combining SaBERT embeddings with multiple recurrent neural architectures and attention mechanisms to understand and classify short texts, resulting in high accuracy. The model effectively captures both semantic and syntactic content in social media comments, allowing it to classify six emotions: Anger, Disgust, Fear, Joy, Sadness, and Surprise. It also includes an additional class named Other for comments that do not clearly express a specific emotion. To evaluate the performance of ApuEmo, we used the TASS 2020 dataset from the Workshop on Semantic Analysis for Task 2: Emotion Detection, alongside a collection of Spanish comments sourced from Facebook related to the Apurimac region in Peru. Furthermore, ApuEmo was compared to the ELiRF-UPV [22] and UMUTeam [23] models, outperforming them in the F1-macro metric by improvements of 0.038 and 0.106, respectively. These results highlight the effectiveness of combining BERT-based word embeddings for the Spanish language with recurrent neural networks and attention layers for emotion classification in short texts.

ApuEmo has added the following novel contributions to be added to existing hybrid methods using pre-trained embeddings with recurrent layers:

- Proposal for a hybrid architecture using SaBERT embedding with recurrent neural network architecture and attention mechanism allowing for better understanding and classification of emotional expressions in social media comments.
- A method of enriching a dataset by combining the TASS 2020 corpus and real comment data from Facebook from participants from the Apurimac region, with more linguistic variability and domain adaptation in Spanish language.
- Evaluation and comparison of hybrid architecture (proposal ApuEmo) through BERT (SaBERT, BETO and RoBERTuito) architecture of pre-trained model in Spanish language, which would show a consistent improvement through its macro F1-score.

The rest of the document is organized as follows. Section II presents the related work. Section III describes the proposed methodology in detail. The results obtained and their respective discussion are described in Section IV. Finally, Section V presents the conclusions and future work.

II. RELATED WORKS

In recent years, the classification of emotions in texts from Spanish social networks has become increasingly important due to the wealth of information available regarding users' opinions, perceptions, and feelings. One approach to this issue is the hybrid model proposed by UMUTeam at the TASS 2020 Workshop on Sentiment Analysis. This model combines linguistic features with statistical methods, allowing for interpretability and competitive results in emotion classification and polarity analysis. It employs Convolutional Neural Networks (CNN) and Support Vector Machines (SVM) with Sequential Minimal Optimization (SMO). However, it does have some limitations concerning scalability and generalization across diverse cultural and dialectical contexts [23]. On the other hand, the TWilBERT model, which is specialized in analyzing Twitter texts in Spanish and was also presented at TASS 2020, has consistently delivered superior results in emotion detection and sentiment analysis. Nonetheless, its optimal performance significantly relies on the amount of domain-specific data that is analyzed [22].

Several studies have used the pre-trained BETO model to classify emotions in social media texts. For example, Bonilla et al. [24] applied this model to the emotional analysis of Facebook comments directed towards media in Costa Rica, identifying predominant emotions such as joy and anger, although limitations were observed due to the size of the dataset and the model's generalization capacity. Similarly, Holgado-Apaza et al. [8] employed BETO to examine emotions in tweets related to COVID-19 in Lima, identifying humour, boredom, and optimism as the most prominent emotions, highlighting the practical utility of emotional analysis during health crises, although limited to specific regional contexts. On the other hand, Rosa and

Chiruzzo [25] identified particular challenges in emotion classification in Spanish, especially for minority classes, proposing to combine multiple corpora to mitigate this limitation, thus emphasizing the need for greater volumes and diversity in annotated data, although they recognized persistent difficulties due to subjectivity in emotional annotation. Additionally, data augmentation techniques using Generative Adversarial Networks (GANs) and Easy Data Augmentation (EDA) have shown significant improvements in emotion classification and sentiment analysis in Spanish, enhancing models like CNN and BETO, although the optimal combination between techniques and specific datasets is yet to be clearly defined [26]. Likewise, Beyond textual modalities, hybrid architectures have also been successfully applied in other domains. For example, EEG-based emotion recognition has been addressed using CNN–LSTM combinations, showing the effectiveness of hybrid deep learning in capturing spatial and temporal dependencies of emotional signals [27]. This further supports our motivation to explore hybrid approaches for textual emotion classification.

Recent studies have emphasized the use of Transformer-based models for emotional classification in social networks in Spanish. The successful implementation of the XLM-RoBERTa model stands out in the EmoEvalEs competition of IberLEF 2021, achieving an average F1 score of 71.70% and demonstrating the ability to handle multilingual texts in Spanish, although with some difficulties in underrepresented emotions [28]. In 2022, the emergence of specialized platforms such as TweetNLP, which provide tools based on Transformers specifically trained for Twitter texts, greatly facilitated emotional analysis and sentiment detection, despite challenges with the informal expressions typical of colloquial Spanish [29]. Simultaneously, other works have explored hybrid methods combining lexical and semantic features with traditional classifiers, showing potential in identifying implicit emotions, although with significant dependence on the quality and size of the lexicon used [30]. Additionally, recent tools like the Pysentimiento library, based on Transformers, have stood out for their effectiveness, accessibility, and versatility in multilingual emotional analysis, though they continue to face challenges related to fairness and biases in specific Spanish contexts [31].

This research contributes to the field of emotional classification in Spanish texts by proposing an innovative hybrid model. This model integrates embeddings generated by SaBERT with an architecture that includes recurrent layers such as Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Gated Recurrent Unit (GRU), and Bidirectional GRU (BiGRU) [32], [33], [34], along with attention layers and fully connected dense layers. The aim of this proposal is to address the limitations previously mentioned, enhancing the accuracy of emotion detection in highly contextual texts derived from social media comments.

III. METHODOLOGY

The Figure 1 shows the architecture of the proposed ApuEmo hybrid model, designed for emotion classification in Spanish. This model is specifically trained using the TASS Task 2 Emotion Detection dataset. The process begins with preprocessing, which includes tokenization and the creation of word embedding representations generated by the SaBERT model—a BERT variant specially trained for the Spanish language. These word embeddings are then input into four different recurrent networks: LSTM, BiLSTM, GRU, and BiGRU. The outputs from these networks are combined and refined through an attention layer that captures the most significant emotional features. Finally, the concatenated representations are processed by a fully connected layer, which classifies the text into one of seven emotions: anger, disgust, fear, joy, sadness, surprise, or others.

A. DATASET

In this section, we describe in detail the datasets utilized for the development of the ApuEmo proposal.

For training and validation purposes, we used the dataset from the TASS 2020 Task 2: Emotion Detection, which was part of the Workshop on Semantic Analysis at SEPLN 2020. This corpus consists of texts in Spanish, extracted from the social media platform Twitter, and it includes seven categories of emotions: Anger, Disgust, Fear, Joy, Sadness, Surprise, and Others. Labels in this dataset were assigned by domain experts following the official annotation guidelines of the TASS competition, ensuring high-quality emotional classification. The distribution of instances per class in the training and validation sets is presented in Table 2.

To compare the proposal with existing models, we used the official test dataset from Task 2: Emotion Detection, which contains 1,665 records.

In addition, we created a complementary dataset consisting of 5,397 comments extracted from Facebook posts using hashtags and keywords related to the Apurimac region in Peru. This collection process was performed using a keyword-based crawler to capture public posts. Since these comments lacked emotion labels, they were manually annotated by a panel of three experts in computational linguistics and NLP (Natural Language Processing), all native Spanish speakers and familiar with emotion taxonomy in the TASS guidelines. To ensure the reliability of manual annotation, Inter-Annotator Agreement (IAA) metrics were calculated. Specifically, Krippendorff's Alpha amongst the three annotators was 0.739, while the average pairwise Cohen's Kappa reached 0.741. According to the Landis and Koch (1977) scale, these values denote substantial agreement, affirming the consistency and quality of emotional labelling within the Facebook corpus.

B. PREPROCESSING

Data preprocessing was designed to minimise textual noise and preserve relevant linguistic features for detecting emo-

TABLE 1. Summary of related works.

Ref.	Authors	Methodology Used	Advantages	Disadvantages	Research Gaps Identified
[28]	Vera et al. (2021)	XLM-RoBERTa on Spanish tweets	High multilingual performance, good overall results	Limited performance in underrepresented emotions	Explore techniques to improve detection of minority emotion classes and data balancing
[29]	Camacho-Collados et al. (2022)	Transformer models specialized in Twitter	Accessible, effective for social media texts	Challenges with informal expressions and idioms	Study contextual handling of informal Spanish expressions in Transformers
[30]	Gonzales-Guerra et al. (2021)	Lexicon-semantic features + supervised models for classification	High interpretability, effective for implicit emotions	Strong dependence on lexicon quality	Develop strategies to dynamically integrate lexical knowledge with neural models
[25]	Rosa and Chiruzzo (2021)	Multiple Spanish corpora combination	Increases emotional data diversity	Subjectivity in annotation and inter-annotator inconsistency	Explore (semi)automatic annotation and cross-validation methods
[31]	Pysentimiento library	HuggingFace Transformer models	Easy to use, multilingual support	Potential biases in specific contexts	Evaluate and mitigate systematic biases across sociocultural variables
[24]	Suárez et al. (2021)	BETO on Facebook comments	Locally adapted, good performance	Small and specific dataset	Expand analysis to multiple countries and cross-cultural comparison
[8]	Holgado-Apaza et al. (2023)	BETO on COVID-19 tweets	Relevant in crisis contexts	Regionally limited scope	Validate model in non-crisis events and other geographic regions
[26]	Data augmentation techniques	GANs, SentiGAN, EDA	Improves baseline model performance	Dataset-dependent effectiveness	Investigate how synthetic augmentation impacts class balance and model stability
[23]	UMUTeam (TASS 2020)	CNN and SVM with linguistic features	High precision, interpretable results	Dialectal variability challenges	Develop semantic representations robust to dialectal and register variation
[22]	TWilBERT (2020)	BERT adapted for Spanish Twitter	Strong performance for Twitter	Requires large domain-specific data	Design lightweight or adaptive models for low-resource contexts

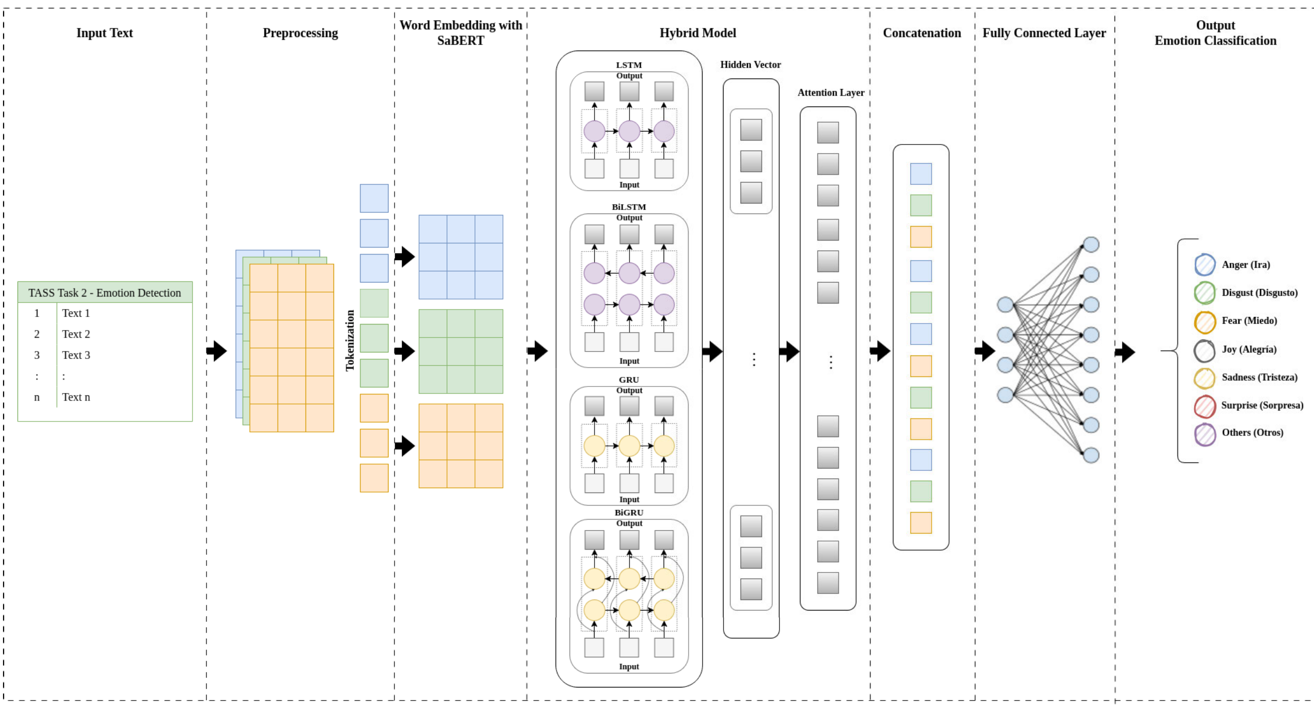


FIGURE 1. ApuEmo architecture: A proposed hybrid architecture based on word embedding with SaBERT and a recurrent neural network.

tions in Spanish. Initially, texts were normalised through the simplification and cleaning of irrelevant tokens such as mentions, URLs, and punctuation marks. Techniques were implemented to remove special characters, including emojis, ASCII emoticons, and symbols foreign to the Spanish alphabet, ensuring that the processed content maintained a structure consistent with the Spanish language. Additionally, anomalous writing patterns were corrected by reducing

TABLE 2. Dataset to train and val from TASS 2020 task 2.

Dataset	Anger	Disgust	Fear	Joy	Sadness	Surprise	Others	Total
Train	600	113	67	1270	706	241	2889	5886
Val	87	16	10	185	103	35	421	857

redundant character sequences, respecting specific spelling rules such as the conservation of "ll" and "rr". Subsequently, orthographic normalisation was performed, which involved removing accents and converting the letter to "n" to ensure compatibility with lexical processing. Semantic representation was enriched through lemmatisation with spaCy, eliminating stopwords except for negation words crucial for sentiment analysis. Finally, emotional labels were normalised and converted into numerical values, preparing the data for training classification models. See more details in Table 3.

TABLE 3. Data cleaning preprocessing techniques.

Technique	Description
Lowercasing	Standardise text casing to avoid discrepancies in lexical analysis.
Removal of unnecessary words	Eliminate generic terms such as "hashtag" and "user" that do not contribute emotional content.
Removal of mentions	Remove direct references to social media users.
Removal of URLs	Delete hyperlinks to prevent bias introduced by external links.
Removal of punctuation marks	Clean the text by removing non-essential orthographic symbols.
Removal of emojis and ASCII emoticons	Eliminate graphical representations and textual emoticons of emotions.
Filtering of non-Spanish characters	Retain only valid characters belonging to the Spanish alphabet.
Orthographic normalisation	Remove diacritics and map "ñ" to "n" to homogenise vocabulary.
Whitespace normalisation	Adjust and standardise spacing between words.
Repeated characters	Correct typographical anomalies while preserving specific linguistic rules.
Lemmatisation and stopword	Reduce words to their base form and remove functional terms, preserving key negation expressions.
Numerical encoding	Convert emotional categories into numerical identifiers.

C. WORD EMBEDDING

To represent the textual inputs in a dense vector space suitable for emotion classification tasks, a transformer-based model specialized in the Spanish language was used. Specifically, SaBERT was employed, a variant of BERT adapted for the semantic representation of sentiments in Spanish, available through the Hugging Face platform.

The embedding generation process was carried out in two levels. Initially, each word was tokenized using the pre-trained SaBERT model, and the corresponding subwords were processed through its internal encoder. Subsequently, the activation vectors for each subword (*last hidden states*) were aggregated by applying an arithmetic average, producing a unique 768-dimensional embedding for each word. This strategy allowed for capturing the semantic information distributed across multiple subwords, thus preserving the contextual representation of each term. To process complete text sequences, each comment was tokenized into individual words, obtaining the corresponding embedding for each word. The sequences were standardized to a maximum length of 50 tokens, with truncation applied to longer sequences, while shorter sequences were filled with null vector padding. In this way, each instance was encoded as a tensor of shape (50, 768), where 50 corresponds to the maximum token length and 768 to the embedding dimension. This ensured a homogeneous representation for subsequent use in the recurrent neural networks designed for emotion classification.

Finally, training and test sets were built from the embedded representations, joining with the numerically coded emotional labels. This embedding methodology not only allowed for the preservation of the semantic context of words but also captured essential emotional nuances for more accurate classification in the Spanish language.

D. HYBRID MODEL

The proposed model for classifying emotions in Spanish texts employs a hybrid approach, integrating multiple Recurrent Neural Networks (RNNs) with attention mechanisms. This design aims to utilize various contextual representations produced by recurrent network variants, capturing both unidirectional and bidirectional temporal dependencies in the data. Initially, embeddings from the SaBERT model are input into four types of recurrent networks: LSTM, BiLSTM, GRU, and BiGRU. Each neural network processes the data sequentially, generating hidden representations that encapsulate diverse linguistic context features.

1) LONG SHORT-TERM MEMORY

The LSTM architecture is designed to mitigate the vanishing gradient problem and preserve relevant information in long sequences. This network employs gating mechanisms to control the flow of information. At each time step t, the forward and backward hidden states are calculated as:

$$\overrightarrow{h}_t = \text{LSTM}(x_t, \overrightarrow{h}_{t-1}) \quad ; \quad \overleftarrow{h}_t = \text{LSTM}(x_t, \overleftarrow{h}_{t+1})$$

where:

- x_t: input vector at time step t.
- $\overrightarrow{h_t}$: hidden state of the address forward in time t.
- $\overleftarrow{h_t}$: hideen state of the address back in time t.
- $\overrightarrow{h_{t-1}}$: previous hidden state (forward direction).
- $\overleftarrow{h_{t+1}}$: hidden future state (backward direction).

These representations allow capturing long-term relationships, such as the interaction between negations and distant emotions in the text.

2) BIDIRECTIONAL LSTM

The BiLSTM extends the LSTM network by processing the sequences in both temporal directions, generating a richer contextual representation. The final representation at each time step is:

$$h_t^{BiLSTM} = [\overrightarrow{h}_t; \overleftarrow{h}_t]$$

where:

- h_t^{BiLSTM}: vector of features concatenated in time t for both directions.
- $\overrightarrow{h_t}$: forward output of the LSTM.
- $\overleftarrow{h_t}$: backward output of the LSTM.

This improves the model's ability to identify emotional patterns distributed across different parts of the text.

3) GATED RECURRENT UNIT

The GRU is a lighter and more efficient alternative to the LSTM. It uses update gates z_t and reset gates r_t to calculate the hidden state h_t at each time step:

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad ; \quad r_t = \sigma(W_r x_t + U_r h_{t-1})$$
$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tanh(W_h x_t + U_h(r_t \odot h_{t-1}))$$

where:

- z_t: update gate vector at time t.
- r_t: reset gate vector at time t.
- W_z, W_r, W_h: weight matrices for the input.
- U_z, U_r, U_h: recurrent weight matrices for the previous hidden state.

This mechanism allows retaining and updating relevant information throughout the sequence with fewer parameters than LSTM.

4) BIDIRECTIONAL GRU

Similar to the BiLSTM, the BiGRU processes the sequence in both directions. The resulting representation is obtained by concatenating the hidden states of both directions:

$$g_t^{BiGRU} = [\overrightarrow{g}_t; \overleftarrow{g}_t]$$

where:

- g_t^{BiGRU}: vector of features concatenated in time t for both directions.
- $\overrightarrow{g_t}$: GRU exit forward.
- $\overleftarrow{g_t}$: GRU exit backwards.

This type of network allows for a more comprehensive understanding of sequential context, especially benefiting tasks like emotion classification.

5) ATTENTION LAYER

After processing the sequence using recurrent networks, an additive attention layer is applied that allows the model to focus on the most relevant parts of the text. Given a set of hidden states $H = \{h_1, h_2, \ldots, h_T\}$, attention is calculated in three steps:

First, an attention score is calculated for each time step:

$$e_t = v^T \tanh(W h_t + b)$$

where:

- h_t is the hidden state at time step t.
- W is the trainable weight matrix that transforms the hidden representation into the attention space and b is the bias vector.
- v is the trainable context vector that determines the relative importance of each hidden state in the sequence.

Then, these scores are normalized using softmax:

$$\alpha_t = \mathrm{softmax}(e_t)$$

Finally, the output is calculated as a weighted sum of the hidden states:

$$h^* = \sum_{t=1}^{T} \alpha_t h_t$$

This mechanism allows the model to focus its attention on the regions of the text with the highest emotional content, generating a more informative and discriminative representation for the final prediction.

The parallel use of LSTM, BiLSTM, GRU, and BiGRU increases the number of parameters and training time compared to a single recurrent architecture. However, this redundancy is deliberate: each variant captures complementary sequential patterns (long-range, bidirectional, and gated dynamics). Their concatenation enhances robustness in imbalanced emotion classes, justifying the trade-off between added complexity and the performance gains reported in the F1-Macro metric (see Section IV).

E. CONCATENATION AND FULLY CONNECTED

1) CONCATENATION

After applying the attention layer to each of the recurrent networks (LSTM, BiLSTM, GRU, and BiGRU), individual contextual vectors are obtained that represent the most relevant information from each architecture. These representations are concatenated to form a single high-dimensional feature vector that groups sequential and semantic information captured from different modeling perspectives.

If h^*_{LSTM}, h^*_{BiLSTM}, h^*_{GRU}, and h^*_{BiGRU} are the output vectors obtained after attention for each network, then the concatenated vector is defined as:

$$h^{concat} = [h^*_{LSTM}; h^*_{BiLSTM}; h^*_{GRU}; h^*_{BiGRU}] \in \mathbb{R}^{d_{concat}}$$

where:

- $h^*_{\mathrm{LSTM}}, h^*_{\mathrm{BiLSTM}}, h^*_{\mathrm{GRU}}, h^*_{\mathrm{BiGRU}}$: output vectors after applying attention to each recurrent network.
- h_{concat}: resulting combined feature vector.
- The concatenated vector $d_{concat} \in \mathbb{R}^{d_{concat}}$, where $d_{concat} = d_{LSTM} + d_{BiLSTM} + d_{GRU} + d_{BiGRU}$: represents the total dimension of the vector after concatenation. This operation allows the complementary combination of features learned by each recurrent submodel.

2) FULLY CONNECTED

The concatenated vector h^{concat} is finally fed into a dense or fully connected layer, which transforms this representation into an output space corresponding to the number of emotional classes. The operation of the fully connected layer is defined as:

$$y = W_{fc} \cdot h^{concat} + b_{fc}$$

where:

- $W_{fc} \in \mathbb{R}^{C \times d_{concat}}$ is the layer weight matrix.
- $b_{fc} \in \mathbb{R}^{C}$ is the bias vector.
- C is the number of emotional classes.
- $y \in \mathbb{R}^{C}$ is the logits vector representing the pre-activation before the softmax function.

F. EMOTION CLASSIFICATION (OUTPUT)

The last stage of the model corresponds to the task of emotional classification, where a label corresponding to one of the previously defined emotional classes is assigned. This stage takes as input the logits vector $y \in \mathbb{R}^{C}$, generated by the fully connected layer, and applies a softmax function to obtain a probability distribution over the emotional classes. Formally, the predicted probability for class i is calculated as:

$$\hat{y}_i = \frac{\exp(y_i)}{\sum_{j=1}^{C} \exp(y_j)} \quad \text{for } i = 1, \ldots, C$$

where C is the total number of emotional classes. In this work, seven classes based on the TASS 2020 Task 2: Emotion Detection dataset are considered: Anger, Disgust, Fear, Joy, Sadness, Surprise, and Others.

Once the probability distribution $\hat{y}$ is obtained, the class with the highest probability is selected as the final label predicted by the model:

$$\text{Predicted Class} = \arg\max_i \hat{y}_i$$

This final prediction is compared with the actual labels during the training process using the cross-entropy loss function, thereby optimizing the model's ability to distinguish between emotions expressed in natural language.

G. ALGORITHM OF PROPOSED HYBRID MODEL

The proposed hybrid model, detailed in Algorithm 1, follows a sequential process beginning with the generation of contextual embeddings using SaBERT. These embeddings are processed in parallel by four recurrent neural architectures LSTM, BiLSTM, GRU, and BiGRU each followed by an attention mechanism to highlight emotionally relevant information. The resulting vectors are concatenated and passed through a fully connected layer to produce class scores, from which the final emotion label is predicted using a softmax function.

H. EXPERIMENTAL SETTINGS

1) HYPERPARAMETER SETTINGS

Table 4 summarises the hyperparameters utilised to train the proposed hybrid model based on recurrent neural networks with attention mechanisms. The model generates contextualised representations using SaBERT and incorporates

Algorithm 1 Proposed Hybrid Model - ApuEmo

Input: Lemmatized token sequence $\{w_1, w_2, \ldots, w_n\}$
Output: Predicted emotion label $\hat{y}$
Generate word embeddings $x_1, x_2, \ldots, x_T$ using SaBERT with subword averaging
Feed embeddings into RNNs:
  (a) LSTM $\rightarrow$ hidden states H_{LSTM}
  (b) BiLSTM $\rightarrow$ hidden states H_{BiLSTM}
  (c) GRU $\rightarrow$ hidden states H_{GRU}
  (d) BiGRU $\rightarrow$ hidden states H_{BiGRU}
for each $H_i \in$ {LSTM, BiLSTM, GRU, BiGRU} **do**
  Apply additive attention to compute $h_i^* = \sum_t \alpha_t^{(i)} \cdot h_t^{(i)}$
end for
Concatenate outputs:
$h^{concat} = [h^*_{LSTM}; h^*_{BiLSTM}; h^*_{GRU}; h^*_{BiGRU}]$
Apply dropout and fully connected layer:
$y = W_{fc} \cdot h^{concat} + b_{fc}$
Apply softmax to y to obtain probabilities $\hat{y}$
Use cross-entropy loss and Adam optimizer to update parameters

various recurrent architectures, including LSTM, BiLSTM, GRU, and BiGRU, each equipped with independent additive attention layers. The training process employs the cross-entropy loss function and the Adam optimiser. The selected hyperparameter values, such as embedding dimension, number of hidden units, and learning rate, were determined empirically based on the performance observed in emotional classification tasks involving Spanish texts.

TABLE 4. Hyperparameter settings for ApuEmo model.

Hyperparameters	Proposed hybrid model (ApuEmo)
Max length	50
Embedding dim	768 (SaBERT)
Attention embedder	Yes (additive attention)
Attention units	Match hidden dim per RNN (e.g., 256, 512)
LSTM units (Layer 1–3)	256
BiLSTM units (Layer 1–3)	256 × 2 (bidirectional)
GRU units (Layer 1–3)	256
BiGRU units (Layer 1–3)	256 × 2 (bidirectional)
Dense units (FC)	Output: 7 (classes)
Optimizer	Adam
Learning rate	0.001
Weight decay	1e-5
Loss function	CrossEntropyLoss
Dropout	0.3
Batch size	16
Epochs	50
Evaluation metric	F1-score (macro)
Hardware	GPU (if available), otherwise CPU

IV. RESULTS AND DISCUSSIONS

A. COMPARISON OF THE PERFORMANCE OF BERT MODELS IN SPANISH

In the Figure 2 and 3, the evolution of the loss value and the F1-Macro metric respectively during the training process of the BETO, RoBERTuito, roBERTaES, BERTin, and SaBERT models is presented. Although a maximum of 50 epochs was defined for training, an early stopping strategy was employed using EarlyStoppingCallback ($early\_stopping\_patience$ = 3), which resulted in all models ending their training at epoch 10. This condition allowed avoiding overfitting and reducing computational time, ensuring that the model maintains a good balance between accuracy and generalization.

In Figure 2, the behavior of the loss function is depicted. All models show a decreasing trend over the epochs, indicating proper convergence during training. RoBERTuito, BETO, and roBERTaES exhibited rapid and sustained reductions, reaching values close to zero towards the tenth epoch, demonstrating good learning ability. Furthermore, SaBERT showed a consistently decreasing loss curve, although with a slightly less pronounced slope. On the other hand, BERTin displayed a persistently high loss with fluctuations throughout the epochs.

On the other hand, the evaluation with F1-Macro in Figure 3 shows that BETO, RoBERTuito, roBERTaES, and SaBERT achieved consistent values from the fourth epoch, stabilizing in a range close to 0.52–0.54. These results indicate robust performance in emotion classification, highlighting the SaBERT and BETO models, whose maximum F1-Macro values reached 0.54. RoBERTuito and roBERTaES also showed a stable evolution of the metric. On the other hand, BERTin obtained the lowest values with a maximum F1-Macro of 0.34, demonstrating limitations in its ability to understand the emotion dataset.

The results obtained reflect a consistent behavior between the Loss and F1-Macro metrics throughout the training. Models such as BETO, RoBERTuito, roBERTaES, and SaBERT exhibited stable loss curves, which translated into competitive performance in the F1-Macro metric, reaching values above 0.52. Regarding the results obtained, SaBERT was selected as the best model, achieving a value of 0.54 in F1-Macro, due to its balance between low loss and high precision in emotion classification.

B. COMPARATIVE ANALYSIS OF SPANISH EMBEDDINGS USING SUPERVISED CLASSIFICATION ALGORITHMS

After evaluating the performance of the embedding models in Spanish, the results showed that BETO, RoBERTuito, and SaBERT achieved higher F1-Macro values and more stable Loss curves, indicating a better generalization ability against imbalanced data compared to roBERTaES and BERTin. For this reason, we used the embedding models generated from BETO, RoBERTuito, and SaBERT to be evaluated using supervised classifiers. To perform the comparison, we used algorithms such as Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), and XGBoost, configured with hyperparameters in each classification algorithm. The analysis of the results is more focused on the F1-Macro metric, which is suitable for imbalanced class classification.

Table 5 presents the performance results of the BETO embedding. With respect to F1-Macro metric, it's observed that the best value obtained was 0.41 using the SVM classifier, while other classifiers such as LR, RF, and XGBoost achieved results of 0.40, 0.28, and 0.36 respectively. These results reflect that, although BETO achieves acceptable accuracies in some classes, its ability for balanced generalization across all classes is limited, affecting overall performance in classification tasks with class imbalance.

The Table 6 shows the performance of the RoBERTuito embedding. The best F1-Macro score achieved was 0.44, achieved with both SVM and XGBoost, surpassing the results of 0.43 in LR and 0.29 in RF. These results indicate that RoBERTuito provides a more suitable representation than BETO for the task of emotion classification under class imbalance, although there is still room for improvement in terms of sensitivity and balanced precision.

The Table 7 presents the results for the SaBERT embedding. The highest F1-Macro score was 0.43, obtained with the SVM and LR classifiers, followed by 0.39 when using XGBoost. These values indicate that the SaBERT embedding, when combined with supervised models, achieves suitable results in emotion classification.

The results obtained in Tables 5, 6, and 7 show that, when analyzing the F1-Macro metric, both RoBERTuito and SaBERT achieved the highest values, standing out over BETO. In particular, RoBERTuito has better results when combined with the SVM classifier with an F1-Macro of 0.44, slightly surpassing other algorithms, while SaBERT obtained competitive but slightly lower values with a difference of 0.01. These results suggest that the RoBERTuito and SaBERT embeddings provide a more effective representation for multiclass emotional classification under class imbalance, especially when used with SVM, demonstrating a better balance between precision and recall across different emotional categories.

C. EVALUATION OF HYBRID MODELS WITH RNN USING EMBEDDINGS IN SPANISH

After evaluating the performance of the embedding models in Spanish, this section explains the performance using hybrid architectures based on RNN. In this stage, the embeddings generated by BETO, RoBERTuito, and SaBERT were used, applying them to models such as LSTM, GRU, BiLSTM, BiGRU, and their combinations, including attention mechanisms. The objective was to determine the impact of each architecture on the task of emotion classification with imbalanced classes. Each configuration was evaluated using the F1-Macro metric, as it provides a more suitable measure in situations of class imbalance. The analysis of the results focused on identifying which combination of RNN model and embedding offered the best balance between precision and recall in multiclass prediction.

The Table 8 shows the results obtained using the BETO embedding in combination with different RNN architectures. It is observed that LSTM and GRU achieved F1-Macro scores of 0.42 and 0.43, respectively. However, the more complex settings integrating multiple layers, such as LSTM + BiL-

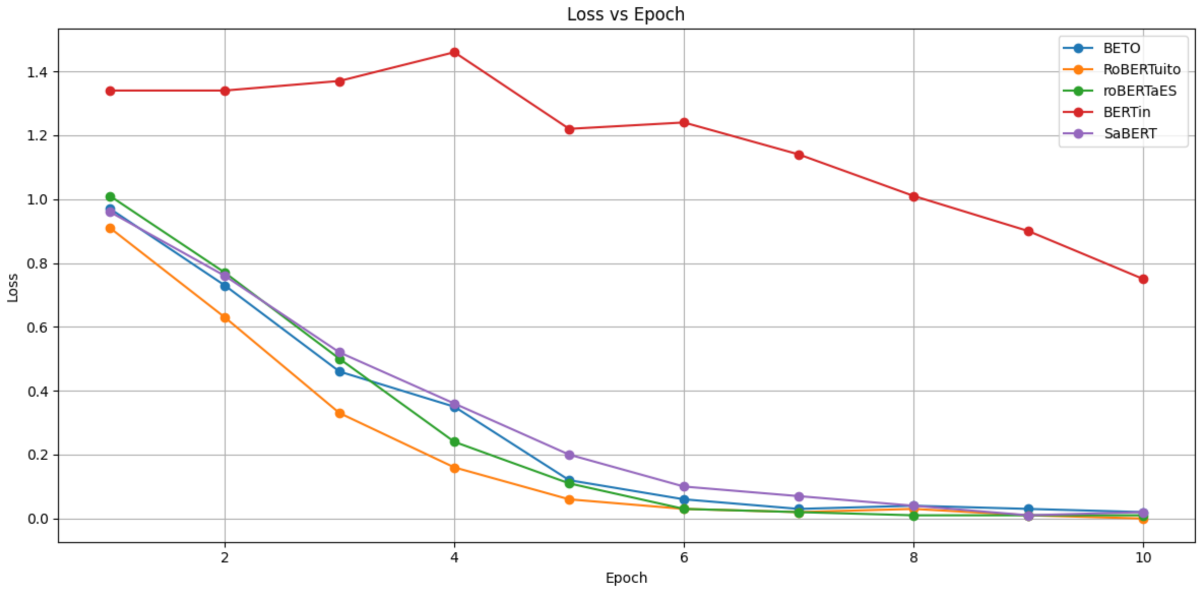


FIGURE 2. Loss function behavior during training of BERT-based models adapted to the Spanish language.

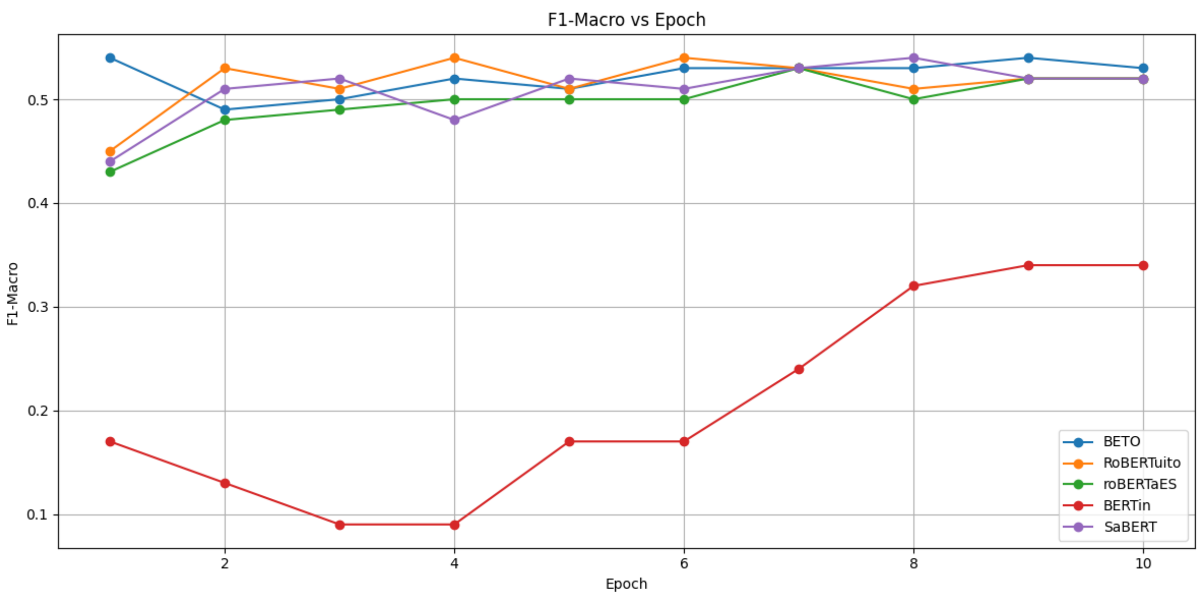


FIGURE 3. F1-macro behavior during training of BERT-based models adapted to the Spanish language.

TABLE 5. Performance metrics using SaBERT embedding with supervised classification models in emotion classification.

Embedding	Classifier	Parameters	Accuracy	Precision	Recall	F1-Macro
BETO	SVM	C = 10, kernel = rbf	0.61	0.54	0.38	0.41
BETO	LR	C = 0.1, penalty = l2	0.61	0.46	0.38	0.40
BETO	RF	max_depth = None, min_samples_split = 2, n_estimators = 50	0.58	0.52	0.25	0.28
BETO	XGBoost	learning_rate = 0.1, max_depth = 3, n_estimators = 100	0.63	0.59	0.33	0.36

STM + GRU + BiGRU and their version with an Attention Mechanism, significantly improved performance, reaching a maximum F1-Macro of 0.47. These results indicate that the use of hybrid architectures and Attention Mechanisms

TABLE 6. Performance metrics using RoBERTuito embedding with supervised classification models in emotion classification.

Embedding	Classifier	Parameters	Accuracy	Precision	Recall	F1-Macro
RoBERTuito	SVM	C = 10, kernel = rbf	0.65	0.52	0.41	0.44
RoBERTuito	LR	C = 0.1, penalty = l2	0.63	0.51	0.41	0.43
RoBERTuito	RF	max_depth = None, min_samples_split = 2, n_estimators = 50	0.58	0.52	0.26	0.29
RoBERTuito	XGBoost	learning_rate = 0.1, max_depth = 3, n_estimators = 100	0.65	0.57	0.35	0.38

TABLE 7. Performance metrics using SaBERT embedding with supervised classification models in emotion classification.

Embedding	Classifier	Parameters	Accuracy	Precision	Recall	F1-Macro
SaBERT	SVM	C = 10, kernel = rbf	0.62	0.49	0.40	0.43
SaBERT	LR	C = 0.1, penalty = l2	0.63	0.50	0.42	0.43
SaBERT	RF	max_depth = None, min_samples_split = 2, n_estimators = 100	0.61	0.53	0.31	0.34
SaBERT	XGBoost	learning_rate = 0.1, max_depth = 3, n_estimators = 100	0.64	0.66	0.36	0.39

TABLE 8. F1-Macro results based on RNN models using BETO embedding.

Embedding	Model	F1-Macro
BETO	LSTM	0.42
BETO	GRU	0.43
BETO	BiLSTM	0.09
BETO	BiGRU	0.35
BETO	BiLSTM + BiGRU	0.37
BETO	BiLSTM + BiGRU + Attention	0.42
BETO	LSTM + BiLSTM + GRU + BiGRU	0.44
BETO	LSTM + BiLSTM + GRU + BiGRU + Attention	0.47

TABLE 9. F1-Macro results based on RNN models using RoBERTuito embedding.

Embedding	Model	F1-Macro
RoBERTuito	LSTM	0.13
RoBERTuito	GRU	0.14
RoBERTuito	BiLSTM	0.09
RoBERTuito	BiGRU	0.14
RoBERTuito	BiLSTM + BiGRU	0.14
RoBERTuito	BiLSTM + BiGRU + Attention	0.16
RoBERTuito	LSTM + BiLSTM + GRU + BiGRU	0.15
RoBERTuito	LSTM + BiLSTM + GRU + BiGRU + Attention	0.14

with BETO embedding has better emotion representation compared to the other individual models that do not use the attention layer.

The Table 9 shows the results of the RoBERTuito embedding applied to RNN models. In this case, the F1-Macro values obtained are considerably lower compared to BETO, ranging between 0.09 and 0.15. Although complex combinations of models were tested, such as LSTM + BiLSTM + GRU + BiGRU + Attention, the maximum recorded F1-Macro was only 0.15. These results suggest that the RoBERTuito embedding, in this context, did not efficiently adapt to the evaluated recurrent architectures, showing limitations in robustly capturing the emotional relationships present in the corpus.

The Table 10 shows results using the SaBERT embedding in combination with various RNN architectures. It can be

TABLE 10. F1-Macro results based on RNN models using SaBERT embedding.

Embedding	Model	F1-Macro
SaBERT	LSTM	0.33
SaBERT	GRU	0.43
SaBERT	BiLSTM	0.03
SaBERT	BiGRU	0.33
SaBERT	BiLSTM + BiGRU	0.32
SaBERT	BiLSTM + BiGRU + Attention	0.41
SaBERT	LSTM + BiLSTM + GRU + BiGRU	0.40
SaBERT	LSTM + BiLSTM + GRU + BiGRU + Attention	0.49

seen that individual models such as LSTM and GRU achieved an F1-Macro of 0.33 and 0.43, respectively. By incorporating combined architectures and attention mechanisms, as in BiLSTM + BiGRU + Attention and LSTM + BiLSTM + GRU + BiGRU + Attention, the performance improved significantly, reaching a maximum F1-Macro of 0.49. This demonstrates that the SaBERT embedding, when enhanced with complex architectures and attention mechanisms, achieves a more balanced and effective emotional representation in multiclass classification scenarios.

In the Tables 8, 9, and 10 we observe that the best overall performance was achieved using the SaBERT embedding combined with the LSTM + BiLSTM + GRU + BiGRU + Attention architecture, reaching an F1-Macro of 0.49. This combination outperformed the others in both precision and classification stability among unbalanced classes. In comparison, although BETO also showed good performance reaching an F1-Macro of 0.47, the value obtained with SaBERT was superior, indicating a better ability to capture complex emotional patterns. On the other hand, RoBERTuito presented considerably lower results, with a maximum F1-Macro of just 0.15, suggesting that its semantic representation was not optimal for tuning in recurrent neural network-based models. These findings highlight the importance of selecting robust embeddings and suitable architectures that maximize the capture of relevant features in emotion and sentiment classification tasks.

TABLE 11. F1-Macro results based hybrid models with different embeddings in emotion classification.

Embedding	Model	F1-Macro
BETO	LSTM + BiLSTM + GRU + BiGRU + Attention	0.47
RoBERTuito	LSTM + BiLSTM + GRU + BiGRU + Attention	0.14
SaBERT	LSTM + BiLSTM + GRU + BiGRU + Attention (ApuEmo proposal)	0.49

TABLE 12. Results of comparing models using TASS 2020 Task 2: Emotion detection dataset.

Model	F1-Macro
ELiRF-UPV [22]	0.45
UMUTeam [23]	0.38
ApuEmo (Proposal)	0.49

In reference to these obtained results (see Table 11), the ApuEmo proposal is based on the hybrid model utilizing SaBERT embedding with LSTM + BiLSTM + GRU + BiGRU + Attention architectures. It demonstrates a greater capacity for emotion representation and classification in the Spanish language compared to other previously evaluated hybrid and individual models.

The Figure 4 presents the confusion matrix show that the model with SaBERT (ApuEmo proposal) achieves the best performance, especially in Joy, Sadness, and Anger, followed by BETO with moderate performance and RoBERTuito with low performance as it concentrates predictions in Others. The F1-Macro values (0.49, 0.47, and 0.14 respectively) confirm the superiority of SaBERT for emotion classification.

D. COMPARISON OF MODELS USING TASS 2020 TASK 2: EMOTION DETECTION

The results of the comparison between two state-of-the-art models and the ApuEmo proposal are presented in Table 12. All were evaluated on the same dataset corresponding to TASS 2020 Task 2: Emotion Detection. The ELiRF-UPV model obtained an F1-Macro score of 0.45, using an adaptation of BERT for tweets in Spanish, followed by a baseline based on Deep Averaging Networks. On the other hand, the UMUTeam model obtained an F1-Macro of 0.38, combining linguistic features with word embeddings and using SVM as a classifier. Finally, the ApuEmo proposal, based on the SaBERT embedding along with a hybrid architecture of LSTM + BiLSTM + GRU + BiGRU integrating attention mechanisms, managed to outperform both models, obtaining an F1-Macro of 0.49, which shows a significant improvement in emotion classification in Spanish texts.

E. COMPARATIVE ANALYSIS OF PREDICTED EMOTIONS

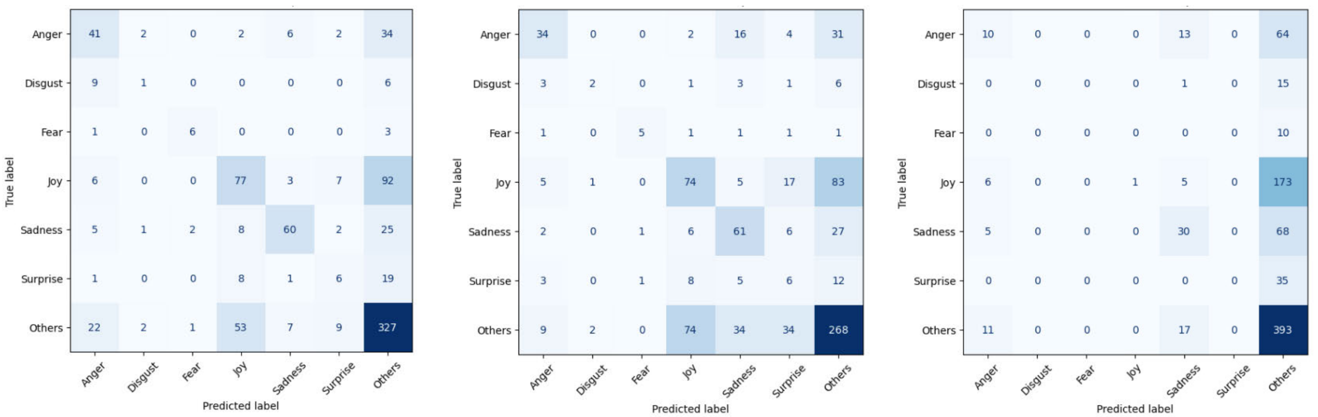
1) PREDICTED EMOTIONS OF TEST TASK 2 DATASET FROM TASS 2020

The Figures 5, 6, 7, and 8 show the distribution of emotions predicted by the models ApuEmo, TweetNLP, PYSentimiento, and XML-RoBERTa, respectively, using the same evaluation dataset (Test Task 2 from TASS 2020). A difference in the behavior of the models with respect to the classifications of different emotional classes is observed. The TweetNLP model (see Figure 6) presents the most balanced distribution among basic emotions, with 498 predictions for Joy, 383 for Sadness, 229 for Disgust, and 14 for Fear, although it maintains a high number of predictions with 507 for Others. This suggests that while the model is capable of identifying multiple emotions, it still tends to group a considerable proportion of instances in the Others class. On the other hand, models like PYSentimiento (see Figure 7) and XML-RoBERTa (see Figure 8) show a greater polarization towards a smaller number of classes, highlighting Joy and Sadness, with scarce representation of classes like Fear or Disgust. This trend could indicate a limitation in the sensitivity of the models to less frequent emotions or those of lower semantic intensity.

In the case of the ApuEmo model (see Figure 5), there is a clear tendency to classify instances into the Others class, with 1,489 predictions, and only limited coverage of Joy and Sadness, with 116 and 32 predictions respectively, and no predictions in Fear and Surprise. Although this behavior may be related to dataset imbalance, it also reflects a conservative approach of the model in the face of semantic ambiguity, prioritizing the Others class when there is not enough contextual evidence. However, considering this result alongside its overall performance (F1-Macro = 0.49), it is inferred that ApuEmo, despite its conservative bias, achieves more accurate classification in the emotions it detects, minimizing false positives in minority classes. This finding reinforces the need to continue developing mechanisms that increase sensitivity towards less represented classes without compromising overall accuracy.

2) PREDICTED EMOTIONS OF FACEBOOK COMMENTS FROM APURIMAC REGION

The Figures 9, 10, 11, and 12 show the distribution of emotions predicted by the ApuEmo, TweetNLP, PYSentimiento, and XML-RoBERTa models, respectively, from Spanish comments extracted from Facebook, related to the ApurÍmac region in Peru. When evaluating the performance of these models, it is observed that the emotion Joy predominates in all predictions, followed by Sadness and Others. This trend could be due to an inherent bias in the dataset or a recurring pattern in automatic emotion classification. The ApuEmo model (see Figure 9) offers broad coverage, recognizing the seven emotions contemplated in the classification, although with a high concentration in the Others class with 2436 predictions, suggesting a conservative strategy in the face of semantic ambiguity. In contrast, TweetNLP (see Figure 10) presents a more balanced distribution among the main emotions, with 1831 predictions for Joy and 1121 for Sadness, while the Others class represents a smaller proportion, with 923 predictions.



SaBERT with LSTM+BiLSTM+GRU+BiGRU+Attention (ApuEmo proposal)

BETO with LSTM+BiLSTM+GRU+BiGRU+Attention

RoBERTuito with LSTM+BiLSTM+GRU+BiGRU+Attention

FIGURE 4. Confusion matrix results based hybrid models with different embeddings in emotion classification.

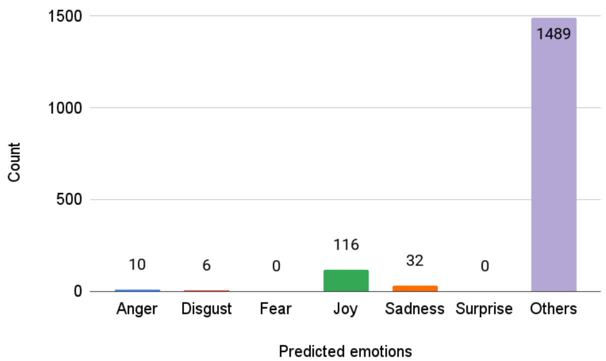


FIGURE 5. Results of emotional classification obtained with the ApuEmo model using Test_task2 dataset.

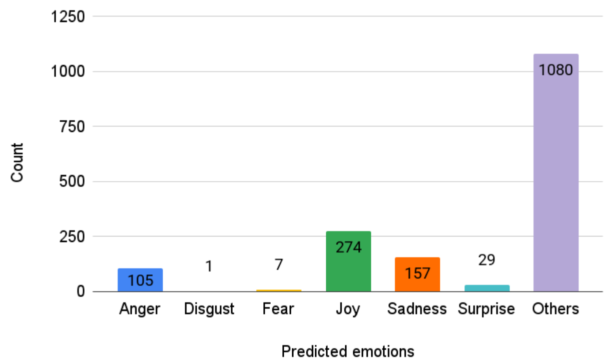


FIGURE 7. Results of emotional classification obtained with the PYsentimiento model using Test_task2 dataset.

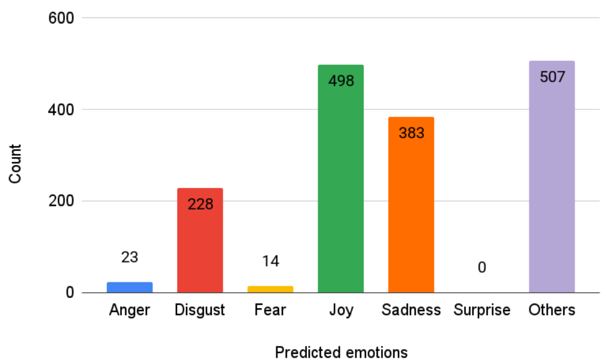


FIGURE 6. Results of emotional classification obtained with the TweetNLP model using Test_task2 dataset.

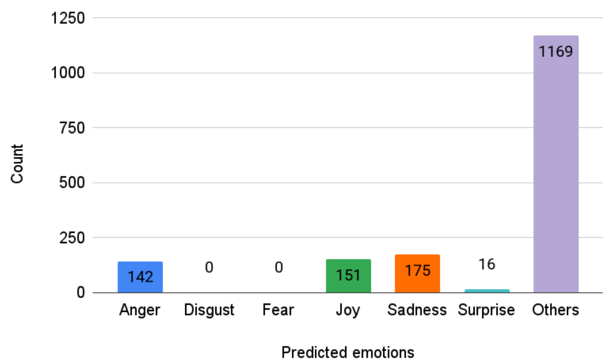


FIGURE 8. Results of emotional classification obtained with the XML-RoBERTa model using Test_task2 dataset.

On the other hand, the PYSentimiento models (see Figure 11) and XML-RoBERTa (see Figure 12) both present limitations in detecting less frequent emotions. PYSentimiento predicts almost exclusively the emotions with 1555 predictions for Joy and 2441 for Others, with scarce representation of other emotions. A similar pattern occurs in XML-RoBERTa, which concentrates most of the predictions with 1657 for Joy and 2207 for Others, while emotions such as Fear, Disgust, and Surprise are not detected at all. These results indicate that although these models have high precision in dominant emotions, their ability to generalize to minority classes is limited, affecting the model's balance. Likewise, ApuEmo stands out for achieving greater diversity in emotional distribution, recognizing all classes of the taxonomic scheme and showing greater sensitivity to emotions with less representation.

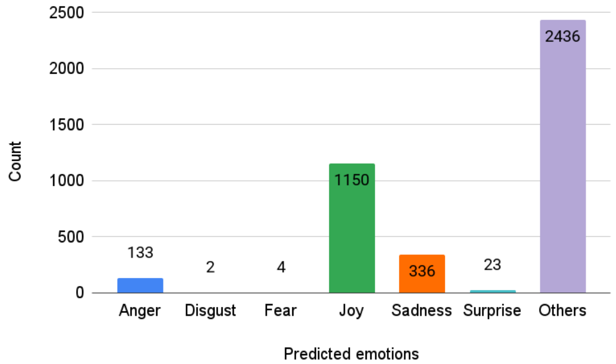


FIGURE 9. **Results of emotional classification obtained with the ApuEmo model using Facebook comments from Apurimac region.**

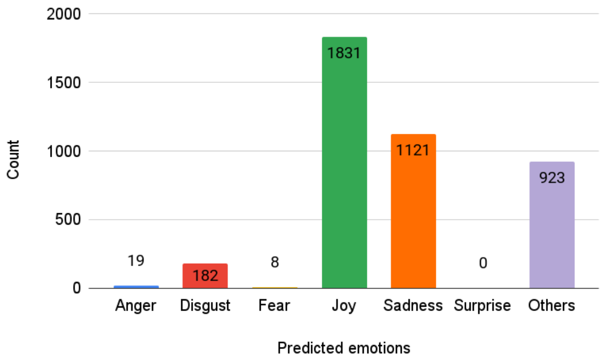


FIGURE 10. **Results of emotional classification obtained with the TweetNLP model using Facebook comments from Apurimac region.**

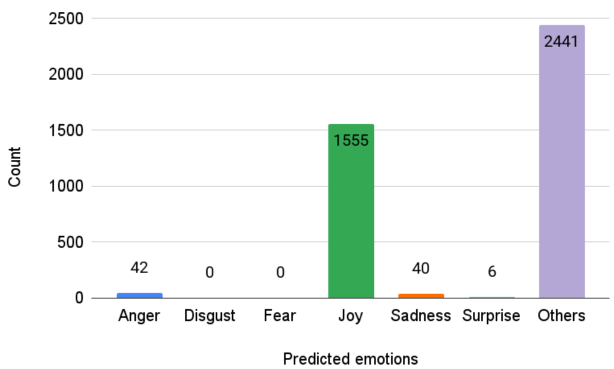


FIGURE 11. **Results of emotional classification obtained with the PYsentimiento model using Facebook comments from Apurimac region.**

Figures 5 to 12 show significant differences in the behavior of the evaluated and compared models when classifying

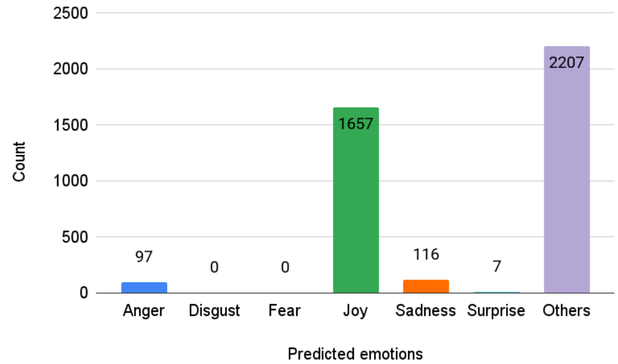


FIGURE 12. **Results of emotional classification obtained with the XML-RoBERTa model using Facebook comments from Apurimac region.**

emotions in the Test Task 2 datasets and comments extracted from Facebook from the Apurimac region, in Peru. In both contexts, a predominance of the Joy class is observed as well as a high concentration of predictions in the Others class, especially using the ApuEmo model and XML-RoBERTa. However, ApuEmo stands out for its ability to recognize the seven emotions in predictions, showing greater coverage and sensitivity towards low-frequency classes such as Fear, Disgust, and Surprise, where other models completely fail. This generalization ability, together with its better performance in terms of F1-Macro, suggests that ApuEmo is the most balanced model for emotion classification tasks in Spanish, applicable to real comments with linguistic variability and noise, as occurs in the extraction of Facebook comments with information from the Apurimac region.

F. FREQUENCY AND LEXICAL POLARITY IN FACEBOOK COMMENTS FROM APURÍMAC IN PERU

The Figure 13 shows a lexical frequency analysis on comments extracted from Facebook related to the Apurimac region in Peru. It is observed that terms like *hermoso, saludos, felicitación, lindo, ApurÍmac* and *Dios* are among the most used, indicating a strong presence of expressions associated with positive emotions and affective connotations. This linguistic pattern reflects a predominant trend in comments considering social appreciation, cultural belonging, and gratitude, which is consistent with the predominance of emotions such as Joy and Others observed in the emotion classification models (see Figures 9, 10, 11, and 12).

The Figure 14 shows a visual representation of the frequently occurring positive and negative words identified in comments extracted from Facebook from the Apurimac region in Peru. In the positive word cloud, terms such as *fiesta, favor, natural, genial, admirable, festival* and *honor*, which are strongly associated with dimensions of joy, respect, and pride, reinforcing the predominance of positive emotions in the overall analysis. On the other hand, the negative word cloud reveals the presence of terms such as *lamentable, horrible, criminal, ruin, error, fatal* and *propaganda*, which are indicative of dissatisfaction and criticism.

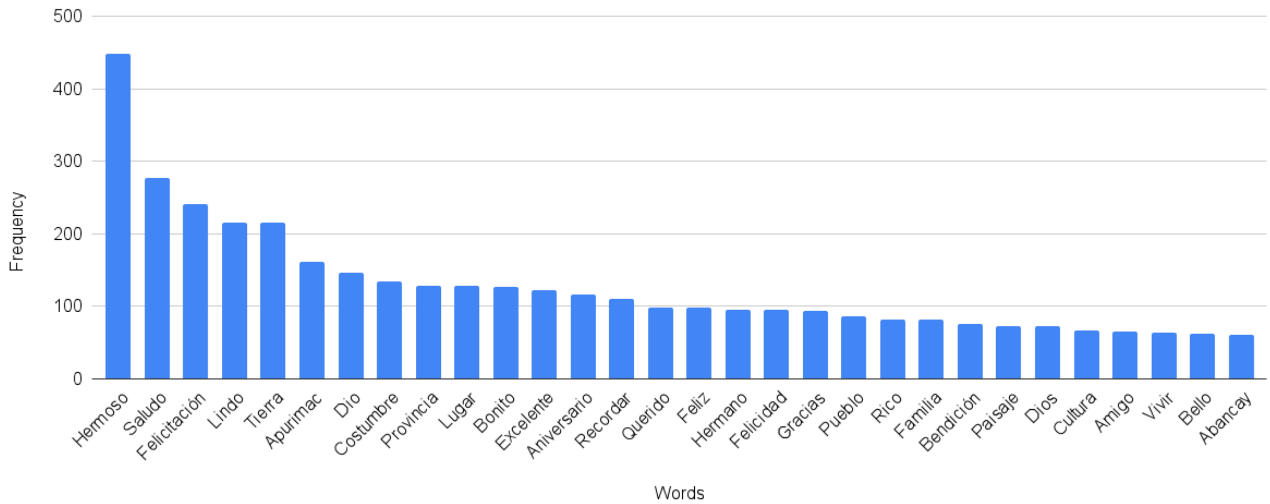
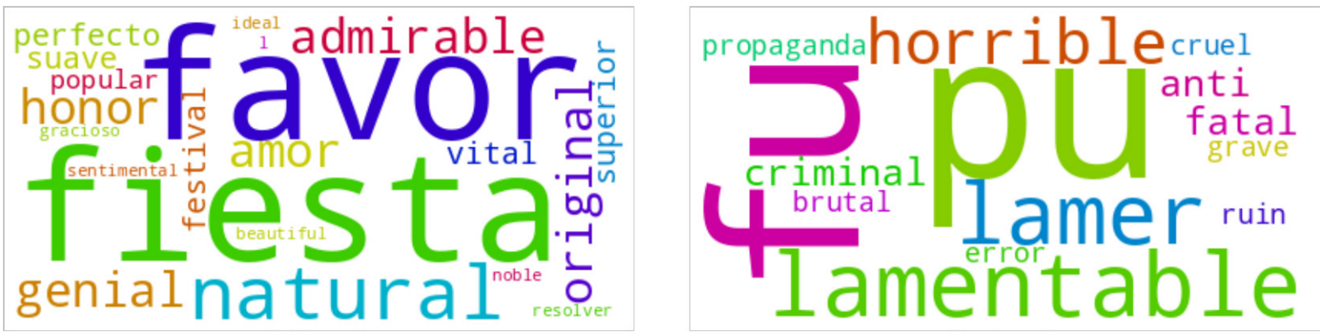


FIGURE 13. Most frequent words in Facebook comments from the Apurimac region in Peru.



Positive words Negative words

FIGURE 14. Word cloud of frequent positive and negative words extracted from Facebook comments from the Apurimac region in Peru.

From a sentiment analysis approach, this polarity differentiation enriches semantic representation models, as it f acilitates the identification of key terms for the adjustment of emotion classification models, particularly when working with noisy comments such as those generated on social networks. Additionally, it shows that even though positive emotions dominate these results, negative emotions are also present and must be considered to achieve precise and balanced detection.

V. CONCLUSION

This study proposed a novel and optimized approach for emotion classification in Spanish comments, leveraging the combination of SaBERT embeddings with recurrent neural networks and attention mechanisms (LSTM, BiLSTM, GRU, BiGRU) and attention mechanisms. Through a rigorous evaluation using the TASS 2020 Task 2: Emotion Detection dataset and comments extracted from Facebook related to the ApurÍmac region in Peru, it was demonstrated that ApuEmo outperforms representative state-of-the-art models, such as ELiRF-UPV and UMUTeam, achieving a maximum F1-Macro value of 0.49. These results demonstrate the model's ability to effectively capture both the semantics and the implicit emotional load in Spanish texts, even in real social contexts characterized by high lexical variability and a high degree of subjectivity.

On the other hand, the experimental results presented in Figures 2 and 3, as well as in Tables 5 to 10, reveal a stable model convergence when using the SaBERT embedding, evidenced by a consistent decrease in the loss function and a progressive improvement in the F1-Macro. These findings validate both the robustness of the training and the model's generalization capability. Regarding the traditional classifiers evaluated with BETO, RoBERTuito, and SaBERT embeddings (Tables 5, 6, and 7), it was observed that the combinations SaBERT + XGBoost and RoBERTuito + SVM offered the best results, although with performance inferior to deep architectures. Also, Tables 8, 9, and 10 show that the best overall performance was achieved through the hybrid LSTM + BiLSTM + GRU + BiGRU + Attention architecture using the SaBERT embedding, reaching an F1-Macro value of 0.49. This configuration not only surpassed traditional models but also equivalent combinations with other embeddings, which reaffirms that integrating recurrent layers and attention mechanisms is key to capturing temporal and semantic relationships in contexts of emotional language in Spanish.

Moreover, complementary lexical and emotional analyses allowed for validating the model's behavior in regional contexts, revealing an emotional distribution consistent with the cultural and linguistic content of the Apurimac region. ApuEmo not only achieved broad coverage of emotions including low-frequency classes such as Fear and Disgust, but also proved robust against the linguistic noise characteristic of social networks. Together, these findings underscore the potential of hybrid models based on language and attention mechanisms for emotional analysis in Spanish, opening new possibilities for its application in public opinion comments, sociopolitical analysis, and decision-making based on collective emotions. As future work, incorporating data balancing strategies such as focal loss or re-sampling could help improve sensitivity in minority emotion classes without compromising overall performance. In addition, future work will apply interpretability techniques by analysing attention weights or visualising token importance, to further validate whether the model emphasises semantically and emotionally meaningful words during classification. Additionally, the exploration of graph-based approaches to enrich the classification of emotions in Spanish texts is proposed. In particular, the integration of architectures like Graph Neural Networks (GNNs), Graph Attention Networks (GATs), or hybrid models that combine semantic embeddings with structured representations from knowledge sources such as WordNet, ConceptNet, or BabelNet is suggested.

REFERENCES

[1] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions* (Studies in Natural Language Processing), 2nd ed., Cambridge, U.K.: Cambridge Univ. Press, 2020.

[2] Y. Mamani-Coaquira and E. Villanueva, "A review on text sentiment analysis with machine learning and deep learning techniques," *IEEE Access*, vol. 12, pp. 193115–193130, 2024.

[3] B. Liu, *Sentiment Analysis and Opinion Mining*, 1st ed., San Rafael, CA, USA: Morgan & Claypool, 2012.

[4] W. G. Parrott, *Emotions in Social Psychology : Essential Readings*. Philadelphia, PA, USA: Psychology press, 2001.

[5] A. de Arriba, M. Oriol, and X. Franch, "Applying sentiment analysis on Spanish tweets using BETO," in *Proc. 37th Int. Conf. Spanish Soc. Natural Lang. Process. Iberian Languages Evaluation Forum (IberLEF)*, 2021, pp. 86–93. [Online]. Available: https://hdl.handle.net/2117/356656

[6] W. Pijal, A. Armijos, J. Llumiquinga, S. Lalvay, S. Allauca, and E. Cuenca, "Spanish pre-trained CaTrBETO model for sentiment classification in Twitter," in *Proc. 3rd Int. Conf. Inf. Syst. Softw. Technol. (ICI2ST)*, Nov. 2022, pp. 93–98.

[7] R. Pan, J. A. García-Díaz, F. Garcia-Sanchez, and R. Valencia-García, "Evaluation of transformer models for financial targeted sentiment analysis in Spanish," *PeerJ Comput. Sci.*, vol. 9, p. e1377, May 2023.

[8] L. A. Holgado-Apaza, C. L. Ancco-Calloapaza, O. Bedregal-Flores, M. Quispe-Layme, and R. Miranda-Castillo, "Measuring the emotional charge: Analysis of the emotions present in the content of tweets about COVID-19 in lima," *Revista Cientfica de Sistemas e Informtica*, vol. 3, p. e587, Jul. 2023.

[9] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics*, vol. 9, no. 3, p. 483, Mar. 2020.

[10] S. Aslan, S. Kızıloluk, and E. Sert, "TSA-CNN-AOA: Twitter sentiment analysis using CNN optimized via arithmetic optimization algorithm," *Neural Comput. Appl.*, vol. 35, no. 14, pp. 10311–10328, May 2023.

[11] Y. Cai, Q. Huang, Z. Lin, J. Xu, Z. Chen, and Q. Li, "Recurrent neural network with pooling operation and attention mechanism for sentiment analysis: A multi-task learning approach," *Knowl.-Based Syst.*, vol. 203, Sep. 2020, Art. no. 105856.

[12] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H.-J. Kang, and J. E. P. Pérez, "Spanish pre-trained BERT model and evaluation data," in *Proc. Workshop Practical ML Developing Countries PML4DC ICLR*. Addis Ababa, Etiopía: International Conference on Learning Representations (ICLR), 2020, pp. 1–9.

[13] A. Fuentes, D. Reinaudo, L. Pardo, and R. Iskandarani. (2023). *Sabert: Spanish Sentiment Analysis Model Based on Bert*. Accessed: Apr. 9, 2025. [Online]. Available: https://huggingface.co/VerificadoProfesional/SaBERT-Spanish-Sentiment-Analysis

[14] J. M. S. Pérez, D. A. Furman, L. A. Alemany, and F. M. Luque, "RoBERTuito: A pre-trained language model for social media text in Spanish," in *Proc. 13th Lang. Resour. Eval. Conf.*, Marseille, France, Jun. 2021, pp. 7235–7243.

[15] J. D. L. Rosa, E. G. Ponferrada, P. Villegas, P. G. D. P. Salas, M. Romero, and M. Grandury, "BERTIN: Efficient pre-training of a Spanish language model using perplexity sampling," *Procesamiento del Lenguaje Natural*, vol. 68, pp. 13–23, Mar. 2022.

[16] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. González-Agirre, C. Armentano-Oller, C. G. Rodriguez-Penagos, and M. Villegas, "MarIA: Spanish language models," *Procesamiento del Lenguaje Natural*, vol. 68, no. 68, pp. 39–60, Mar. 2022, doi: 10.26342/2022-68-3.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," Tech. Rep., 2018.

[18] H. Zhang and M. O. Shafiq, "Survey of transformers and towards ensemble learning using transformers for natural language processing," *J. Big Data*, vol. 11, no. 1, p. 25, Feb. 2024.

[19] R. Vatambeti, S. V. Mantena, K. V. D. Kiran, M. Manohar, and C. Manjunath, "Twitter sentiment analysis on online food services based on elephant herd optimization with hybrid deep learning technique," *Cluster Comput.*, vol. 27, no. 1, pp. 655–671, Feb. 2024.

[20] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Hybrid models for emotion classification and sentiment analysis in Indonesian language," *Appl. Comput. Intell. Soft Comput.*, vol. 2024, no. 1, Jan. 2024, Art. no. 2826773.

[21] K. W. Tan, W. N. Lim, and Y. Lee, "Emotion classification through text with proposed hybrid learning emotion model," in *Proc. 3rd Int. Conf. Artif. Intell., Inf. Process. Cloud Comput.*, Jun. 2022, pp. 1–5.

[22] J. Á. González, J. A. Moncho, L.-F. Hurtado, and F. Pla, "ELiRF-UPV at TASS 2020: TWilBERT for sentiment analysis and emotion detection in Spanish Tweets," in *Proc. Iberian Lang. Eval. Forum (IberLEF)*, 2020, pp. 179–186.

[23] J. A. García-Díaz, Á. Almela, and R. Valencia-García, "UMUTeam at TASS 2020: Combining linguistic features and machine-learning models for sentiment classification," in *Proc. Iberian Lang. Eval. Forum (IberLEF)*, 2020, pp. 187–196.

[24] F. D. S. Bonilla, N. V. Oconitrillo, J. D. G. Castro, A. Durán, and A. P. J. Chavarría, "BETO emotion analysis of Facebook users reacting to major media outlets in Costa Rica," in *Proc. 3rd Int. Conf. Transdisciplinary AI*, Sep. 2021, pp. 107–111.

[25] A. Rosá and L. Chiruzzo, "Emotion classification in Spanish: Exploring the hard classes," *Information*, vol. 12, no. 11, p. 438, Oct. 2021.

[26] R. Gutiérrez Benítez, A. Segura Navarrete, C. Vidal-Castro, and C. Martínez-Araneda, "Guide for the application of the data augmentation approach on sets of texts in Spanish for sentiment and emotion analysis," *PLoS ONE*, vol. 19, no. 9, pp. 1–24, Sep. 2024.

[27] B. Chakravarthi, S.-C. Ng, M. R. Ezilarasan, and M.-F. Leung, "EEG-based emotion recognition using hybrid CNN and LSTM classification," *Frontiers Comput. Neurosci.*, vol. 16, Oct. 2022.

[28] D. Vera, Ó. Araque, and C. Á. Iglesias, "GSI-UPM at IberLEF2021: Emotion analysis of Spanish tweets by fine-tuning the XLM-RoBERTa language model," in *Proc. Iberian Lang. Eval. Forum (IberLEF)*, Málaga, Spain, 2021, pp. 16–26.

[29] J. Camacho-Collados, K. Rezaee, T. Riahi, A. Ushio, D. Loureiro, D. Antypas, J. Boisson, L. Espinosa Anke, F. Liu, and E. Martínez Cámara, "TweetNLP: Cutting-edge natural language processing for social media," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, W. Che and E. Shutova, Eds., Abu Dhabi, UAE: Association for Computational Linguistics, 2022, pp. 38–49, doi: 10.18653/v1/2022.emnlp-demos.5.

[30] H. Gonzalez-Guerra, A. Simón-Cuevas, J. M. Perea Ortega, and J. A. Olivas, "Un enfoque semántico en la selección de características basadas en léxico para la detección de emociones," *Procesamiento del Lenguaje Natural*, vol. 67, no. 67, pp. 115–126, 2021. [Online]. Available: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6382

[31] J. M. Pérez, M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, and M. V. Martínez, "Pysentimiento: A Python toolkit for opinion mining and social NLP tasks," 2024, *arXiv:2106.09462*.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[33] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[34] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

YONATAN MAMANI-COAQUIRA received the bachelor's degree in computer and systems engineering from the National University Micaela Bastidas of Apurimac, Peru, in 2013, and the master's degree in computer science from the National University of San Antonio Abad of Cusco, Peru, in 2021. He is currently pursuing the Ph.D. degree in engineering, with a specialization in computer science with the Pontifical Catholic University of Peru (PUCP).

He is a conducting Researcher with the Artificial Intelligence Group, PUCP. He previously had an internship and collaborated as a Researcher with the Data Analytics Science and Engineering Laboratory, Pablo de Olavide University, Seville, Spain. He is currently a Professor with the Faculty of Engineering, National University Micaela Bastidas of Apurimac. His research interests include machine learning, deep learning, data analytics, computer vision, and natural language processing.

MANUEL J. IBARRA-CABRERA received the bachelor's degree in computer and systems engineering from the National University San Antonio Abad of Cusco, Peru, in 2001, the master's degree in computer science from the University of Chile, in 2012, and the Ph.D. degree in computer science from the National University of Altiplano of Puno, Peru, in 2019.

He also completed internships at the Department of Industrial Design, Eindhoven University of Technology, The Netherlands, the National University of San Marcos, Peru, and the University of Valencia, Spain. He is currently a Professor and a Researcher with the National University Micaela Bastidas of Apurimac, Peru. His research interests include software engineering, artificial intelligence, educational games, mobile computing, and collaborative systems.

MARIO AQUINO-CRUZ received the bachelor's degree in systems and computer engineering from the Technological University of the Andes, Peru, in 2004, and the master's degree in computer science from the National University of the Altiplano of Puno, Peru, in 2020, where he is currently pursuing the Ph.D. degree in statistics and computer science.

He is an Active Researcher with the Artificial Intelligence and its Applications Group, National University Micaela Bastidas of Apurimac, Peru. He is currently a Professor with the Faculty of Engineering, National University Micaela Bastidas of Apurimac. He has participated as a Lecturer in academic and scientific events in Brazil, Ecuador, Mexico, Chile, USA, and Spain. His research interests include machine learning, deep learning, computer vision, and natural language processing.

WILSON J. MOLLOCONDO-FLORES received the bachelor's degree in statistics science and the master's degree in computer science from the National University of the Altiplano of Puno, Peru, in 1996 and 2007, respectively, and the Ph.D. degree in administration from the University of Alas Peruanas, Peru, in 2013.

He is conducting research on dynamic testing through R-Studio and Sweave, and carries out studies in business and social sciences, applying techniques such as machine learning, multivariate analysis, and structural equation modeling. He is currently a Senior Lecturer with the Faculty of Engineering, National University Micaela Bastidas of Apurimac, Peru.