ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.273

Use of Bi-LSTM for Emotion Detection via Text Messaging in the Latin American Context

Alejandro Jose Junior Cruz Mocarro[1]; Brayan Smith Pariona Castillo[2]; Ruben Oscar Cerda Garcia[3]
[1,3]*Peruvian University of Applied Sciences, Perú, U202114593@upc.edu.pe, pcsircer@upc.edu.pe*
[2]Peruvian University of Applied Sciences, Perú, *U20211C374@upc.edu.pe*

Abstract*- Emotional violence in digital environments represents a critical and growing challenge in Latin America, where regional slang, cultural expressions, and informal speech patterns make automated detection particularly difficult. Studies from Peru indicate alarming rates of domestic violence among adolescents and suicidal ideation in adult populations, underscoring the need for context-sensitive and locally adapted risk-detection tools. Most existing natural language processing models are trained on English or neutral Spanish corpora, limiting their ability to interpret the communicative nuances of local informal language. To address this gap, this study proposes a Bidirectional Long Short-Term Memory network trained on an 18,000-message dataset collected via automated web scraping from Twitter, capturing authentic Latin American informal speech. The dataset was built using a semi-automatic two-stage labeling strategy combining a pre-labeled English violence corpus translated into Spanish with keyword-defined classification of regional tweets. A balanced 50/50 class distribution was maintained across both training and validation splits to prevent any model bias. The complete pipeline spans data collection and cleaning through translation and binary classification of violent versus non-violent messages. The model achieves 90 percent accuracy and an F1-score of 0.89, outperforming standard LSTM and GRU baselines trained under identical conditions. It demonstrates semantic understanding by correctly classifying ambiguous regional expressions, distinguishing affectionate uses of violent language from genuine threats. Unlike keyword-based approaches, the model captures contextual intent rather than surface-level patterns, reducing false positives in everyday colloquial speech. Ethical safeguards including data anonymization and compliance with platform terms of service were applied throughout the research process.

***Keywords:* Natural Language Processing; Bidirectional LSTM; Emotion Detection; Violent Language; Latin American Spanish.**

I. INTRODUCTION

Domestic violence and suicidal ideation are two critical public health crises in Peru and Latin America that severely impact mental health. At a health center in Pichari, Cusco, an elevated risk of domestic violence was reported among adolescents and young women, with higher risk among females across all life stages [1]. This exposure deteriorates emotional health and demands urgent public policy responses. In parallel, in Lambayeque, Peru, 28.3% of adults were found to have suicidal ideation during the COVID-19 pandemic, with higher prevalence among women, widowed individuals, low-income groups, and those who experienced family losses [2]. These data underscore the need to strengthen emotion detection and prevention systems tailored to the Latin American sociocultural context.

Automated emotion detection in text can be pivotal for identifying risk situations that currently go unnoticed. In contexts such as depression or cyberbullying, it enables the activation of early alerts and timely interventions [3]. Moreover, NLP models have demonstrated effectiveness in detecting violent language, reinforcing their applicability to safer digital environments [4]. Their adaptable architecture also makes it possible to scale this solution to other linguistic variants.

Despite the potential impact of emotion-detection systems, their development in contexts such as Peru faces significant challenges. One is the scarcity of representative data, since most models (GRU, BERT, RoBERTa, etc.) are trained on corpora in English or neutral Spanish, which limits their adaptation to specific dialectal variants and cultural realities [5]. In addition, the complexity of emotional language—including sarcasm, slang, and ambiguity—hampers precise interpretation, especially in non-standardized languages [6]. Compounding this, many structured datasets used to train and evaluate artificial intelligence models do not reflect communicative patterns, introducing bias and limiting model reliability in critical situations.
The problem persists due to methodological limitations that have impeded significant progress: traditional approaches—based on keywords or shallow analysis—do not capture the emotional complexity of language, especially in cases such as sarcasm or ambivalence, thereby undermining accuracy in real-world applications [7].

The development of the proposal is divided into two major steps: first, the use of web scraping on Twitter to feed the Bi-LSTM model, prioritizing the detection of violent patterns and, with that, negative emotions; and second, the implementation of the Bi-LSTM model to analyze the collected data and determine whether a message represents a potential risk. Among the project's limitations are the need to continually update the model to adapt to new forms of communication that may pose a danger to the user. In addition, the system may struggle to interpret sarcasm or highly contextual language.

II. RELATED WORKS

In their research, [8] propose improving emotion detection in texts about animal testing by applying deep learning techniques. The study uses an LSTM model together with Word2Vec to vectorize words and classify sentiments in more than 15,000 tweets collected through the Twitter API. The proposal aims to overcome the limitations of correctly interpreting complex emotions on social networks. As a result, the model achieved an accuracy of 88.7% and an F1-

score of 0.87; however, the absence of a neutral category was identified as a limitation, as well as the possibility of obtaining better results by using more sophisticated models.

For their part, [9] present a hybrid model composed of SemGloVe, BERT, and Bi-LSTM to improve question–answer classification in the field of natural language processing. This approach seeks to overcome the semantic limitations of earlier techniques such as GloVe, enabling more accurate contextual understanding. The model was trained using the CR23K and CR100K datasets, reaching an accuracy of 92% and outperforming approaches such as GPT-2 and SPARQL. Despite these promising results, the authors highlight the importance of addressing data imbalance and exploring alternatives such as RoBERTa in future research.

Similarly, [10] propose a hybrid model that combines CNN, Bi-LSTM, and a CTC decoder to improve handwritten text recognition. This approach extracts both spatial and sequential features from images, overcoming limitations of traditional OCR systems. The model was evaluated on the IAM and RIMES datasets, achieving accuracies of 98.55% and 98.80%, respectively, and showing minimal word transcription errors. Nonetheless, the study acknowledges as limitations its lower performance on fragmented texts and its limited adaptability to different languages or contexts.

In a different approach, [11] presents a model that combines SNN and ConvLSTM to optimize English vocabulary learning among non-native speakers. The proposal seeks to overcome the limitations of traditional methods through an architecture adaptable to different corpus sizes and linguistic structures. Trained with data from WordNet and the Oxford English Corpus, the model achieved accuracy of up to 84% and an F1-score above 78%, depending on the corpus size used. It should be noted as a limitation that its use is restricted solely to the English language.

Complementarily, [12] propose a hybrid model that combines LSTM, Bi-LSTM, and logistic regression with an attention mechanism to improve the detection of Arabic dialects on social networks. The study addresses the challenge of identifying dialects due to their linguistic variations and the lack of standardized resources. Using 34,905 comments extracted from Twitter and techniques such as Word2Vec with Skip-gram, the model achieved an accuracy of 88.73%, surpassing previous methods such as MARBERT and AraBERT. Conversely, the lack of evaluation on other dialects is noted as a limitation, which could affect its capacity for generalization.

III. BI-LSTM MODEL FOR EMOTION DETECTION

This section describes the main contribution of the study. As a solution, we propose an approach that combines web scraping techniques to collect real-world data from Twitter, an NLP model for emotion detection, and a user-adapted alert interface.

A. Preliminary Concepts

Definition 1 (Bi-LSTM [13]): A type of recurrent neural network (RNN) that processes input sequences in both directions—forward and backward—allowing it to capture the preceding and following context of each word, which improves emotional or contextual understanding in texts.

Definition 2 (Emotion detection [14]): The process of automatically identifying emotional states (such as anger, sadness, or fear) from textual data using Natural Language Processing (NLP) and Machine Learning techniques.

Definition 3 (Data preprocessing [15]): The set of techniques used to clean, normalize, and transform raw data to improve the performance of machine learning models. It includes tasks such as emoji removal, text correction, or content translation.

Definition 4 (Tokenization [16]): The process of splitting a text into smaller linguistic units called tokens—which may be words, phrases, or characters—that serve as the basis for language analysis and modeling.

Example: Splitting the sentence “No estoy bien” into the tokens ["No", "estoy", "bien"].

Definition 5 (Corpus [17]): A structured collection of texts compiled for linguistic or computational purposes, used to train, evaluate, or analyze natural language models.

Definition 6 (Web scraping [18]): An automated method for extracting information from websites, commonly used to collect user-generated content on platforms such as Twitter, in order to build datasets for training models.

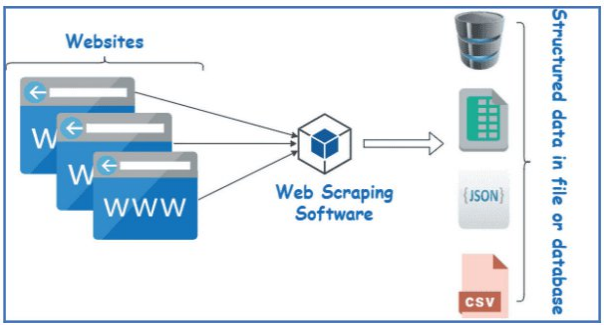

Fig. 1 Basic Flow – Web Scrapping [19].

Definition 7 (API [20]): An Application Programming Interface (API) is a set of functions and protocols that enable communication between different systems or applications. In text-processing contexts, APIs are used to access messaging platforms, collect data, or integrate external services.

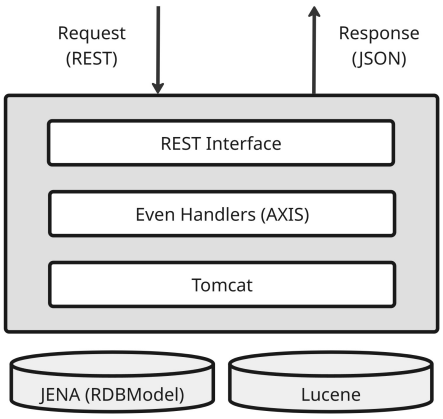

Fig. 2 API Flow.

B. Background and Development

The proposed model is structured in two phases: **Tweet Harvest** (data extraction) and **Model Learning and Validation** (training and validation). This end-to-end approach enables emotion detection in text within the Latin American linguistic and cultural context. Given the rise of emotional violence in digital environments, we prioritize real-world data that reflect everyday language, including regional expressions where violence may be normalized. The proposal seeks not only a technical solution but also a preventive tool for scenarios of emotional risk.

Bi-LSTM Model

Figure 3 shows the Bi-LSTM model, which processes information along two parallel paths: one in the forward direction (forward layer) and the other in the reverse direction (backward layer). Each LSTM layer captures patterns from different temporal perspectives, and their outputs are combined to feed an activation layer, which ultimately produces the corresponding output. This structure makes it possible to capture complex relationships in the data and provide more contextualized interpretations in sequential modeling tasks.

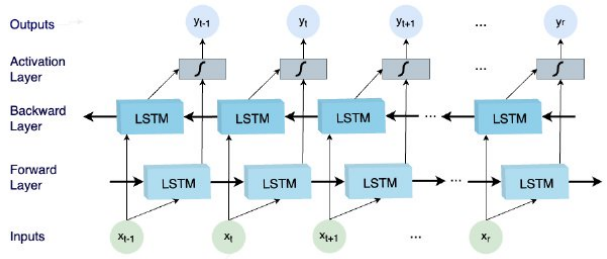

Fig. 3 Bi-LSTM architecture [21].

Tweet Harvest

The Tweet Harvest stage begins with TweetExtraction, which collects posts from Twitter via web scraping. Then, TweetCleaning normalizes the text and removes non-textual elements such as emojis, and TweetTranslate automatically translates English messages into Spanish. The entire process is automated with Python scripts, generating a labeled dataset (training dataset) ready to train the Bi-LSTM model. Unlike approaches that use neutral or international datasets, this solution focuses on messages with emotional content from the Latin American context, including both violent and non-violent expressions, based on slang and linguistic patterns characteristic of the region's informal Spanish.

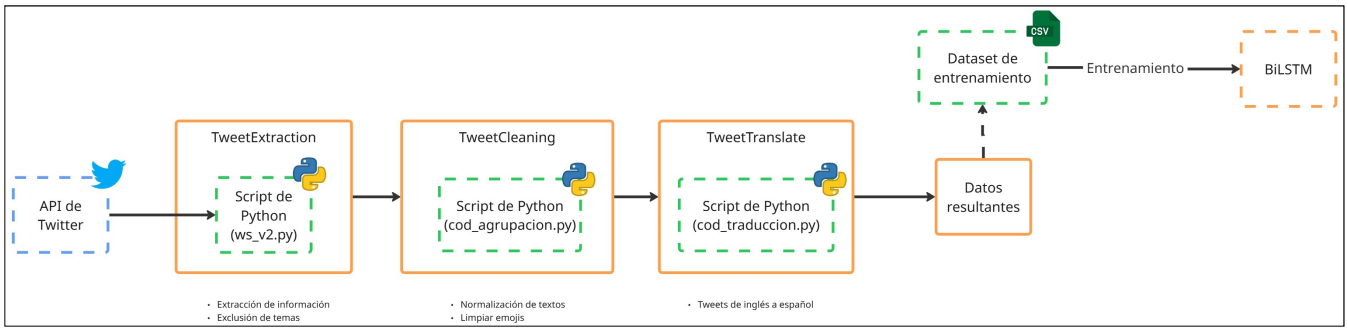

Fig. 4 ETL diagram for the Bi-LSTM model's training database.

On the other hand, the contents and operation of the scripts are as follows:

TweetExtraction Script (ws_v2.py)

A Python script was developed to automate the collection of tweets and classify them according to positive or negative emotions, using expressions typical of both general and regional (Latin American) speech. The process includes cleaning, semantic filtering, topic exclusion (such as soccer or reality shows), and storage in Excel.

Key expressions are defined and grouped by emotion type (e.g., threats, affection, friendship). These serve to identify the type of content in the collected texts.

This routine queries the Twitter API to obtain recent Spanish-language posts that contain key expressions related to a specific emotion, applying a filtering system to improve the relevance of results and avoid exceeding API usage limits.

Each tweet is cleaned, classified, and enriched with metadata such as date, user, location, likes, retweets, and hashtags. It is retained only if it meets the emotional criteria.

Data collection followed Twitter's Terms of Service using publicly available posts only. No personally identifiable information was stored, and the dataset was anonymized prior to model training to ensure user privacy compliance.

Fig. 5 Twitter Data Extraction Flow.

NOMBRE_COLUMNA	DESCRIPTIVO	VALOR_EJEMPLO
ID	Identificador único de mensaje	0022DWKP
Fecha	Fecha del Tweet	2025-04-16 20:16:24
Usuario	Usuario que publicó tweet	Kary_ntvg
Texto_Original	Mensaje original traído de Twitter	yeah he raped me #RAPED
Tipo_Emocion	Emoción vinculada al Tweet	negativa_violenta
Categoria_Emocion	Categoría del Tweet de acuerdo al contenido	sexual_violence

Fig. 6 Descriptive Excel table resulting from the first script.

TweetCleaning Script (cod_agrupacion.py)

This Python script automates the aggregation of messages from multiple Excel files, performs cleaning on the tweets, and finally exports all results into a single Excel file. It also uses the pandas library for data manipulation and tqdm to display progress bars.

This routine performs basic cleaning of tweet text prior to analysis or classification. Its goal is to remove irrelevant elements or those that could introduce noise in downstream processing—such as visual expressions or special characters—ensuring the text is more uniform and useful for NLP models or manual analysis.

Fig. 7 Tweet Cleaning Workflow.

NOMBRE_COLUMNA	DESCRIPTIVO	VALOR_EJEMPLO
ID	Identificador único de mensaje	0022DWKP
MENSAJE_ORIGINAL	Mensaje original traído de Twitter	yeah he raped me
TIPO_VIOLENCIA	Tipo de violencia del mensaje	sexual_violence

Fig. 8 Descriptive Excel table resulting from the second script.

TweetTranslate Script (cod_traduccion.py)

This script was designed to automatically translate messages written in English into Spanish, starting from an Excel file containing previously collected messages. Translation is performed only when needed—that is, if the detected language of the message is English. It uses the googletrans library, pandas for data manipulation, and tqdm to show real-time progress.

The traducir_mensaje function automatically detects the language of the text via translator.detect. If it identifies English ('en'), it translates the text into Spanish with translator.translate; otherwise, it preserves the original message. It returns the final text and a binary flag (1 if translated, 0 if not). In addition, it updates the tqdm progress bar with the process status: translated, not required, or error.

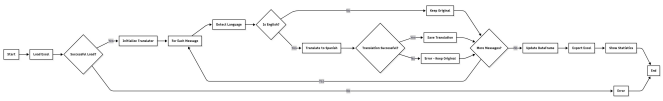

Fig. 9 Tweet Translation Workflow.

NOMBRE_COLUMNA	DESCRIPTIVO	VALOR_EJEMPLO
ID	Identificador único de mensaje	0022DWKP
MENSAJE_ORIGINAL	Mensaje original traído de Twitter	yeah he raped me
TIPO_VIOLENCIA	Tipo de violencia del mensaje	sexual_violence
MENSAJE_TRADUCIDO	Mensaje original traducido al español	sí me violó
FLAG_TRADUCCION	Flag que indica si el mensaje fue traducido	1

Fig. 10 Descriptive Excel table resulting from the third script.

Model Learning and Validation

The first process, Model Learning, which consists of training the Bi-LSTM model, is carried out in Google Colab, leveraging its cloud execution environment and direct integration with Google Drive for file uploads. The model is trained on the textual dataset generated through the Tweet Harvest process, which yields a dataset containing 18,000 messages labeled as violent or non-violent. Approximately 20% of the total—around 3,000 messages—is used for the Model Validation process.

Dataset labeling followed a two-stage approach. First, English-language messages were sourced from a pre-labeled violence corpus containing categories such as physical violence, emotional violence, and sexual violence, which were unified into a single 'violent' class and translated into Spanish via the TweetTranslate script. Second, Spanish-language tweets collected via TweetExtraction were automatically classified using manually defined keyword groups characteristic of Latin American informal speech — such as explicit threats, affective expressions, and regional insults — and reviewed for consistency. This semi-automatic labeling strategy ensured both scalability and contextual relevance for the Latin American linguistic context.

Table I presents representative samples from the corpus, illustrating how regional slang and colloquial expressions were labeled to reflect the linguistic particularities of the Latin American context.

TABLE I
REPRESENTATIVE EXAMPLES FROM THE TRAINING CORPUS ILLUSTRATING PERUVIAN AND LATIN AMERICAN INFORMAL SPANISH

#	Expression (Spanish)	English Translation	Label
1	Te voy a matar a besos, cuidado.	I'm going to kill you with kisses, watch out.	Non-violent (0)
2	Estuvo bien bacán el concierto, ojalá vayamos juntos.	The concert was really cool, hope we go together.	Non-violent (0)
3	Hoy estuvo chévere la reunión, ojalá se repita.	The meeting was great today, hope it happens again.	Non-violent (0)
4	Oe imbécil, cállate ya antes que te reviente.	Hey idiot, shut up before I smash you.	Violent (1)
5	No me vengas con huevadas, causa, porque te saco la mierda.	Don't come at me with nonsense, bro, or I'll beat you up.	Violent (1)
6	Eres una inútil, no sirves para nada.	You're useless, you're good for nothing.	Violent (1)

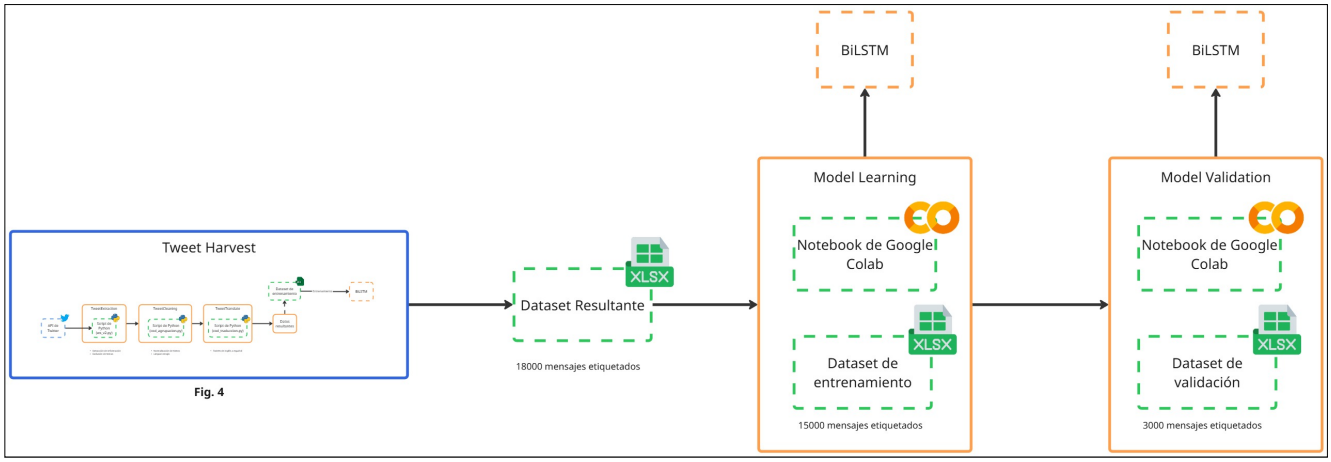

Fig. 11 Bi-LSTM model training process

Below is the Model Learning workflow, divided by notebook blocks:

Fig. 12 Data Loading cell.

This cell loads Excel files from Google Drive, separating the training set (df_train) and the validation set (df_test). These files contain two key columns: Texto (the message) and Violento (the label: 0 for non-violent, 1 for violent).

```
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
import numpy as np

# Entrenamiento
texts_train = df_train['Texto'].astype(str).tolist()
labels_train = df_train['Violento'].tolist()
```

Fig. 13 Text Extraction cell.

Texts and labels are extracted and converted into lists, repeating the step for the validation data. A Keras Tokenizer is then created to convert words into integers, and it is fit only on the training data.

```
# Tokenizador solo se entrena con los datos de entrenamiento
tokenizer = Tokenizer(num_words=10000, oov_token='<OOV>')
tokenizer.fit_on_texts(texts_train)
```

Fig. 14 Tokenizer Creation cell.

```
# Convertir a secuencias
sequences_train = tokenizer.texts_to_sequences(texts_train)
sequences_test = tokenizer.texts_to_sequences(texts_test)

# Padding
padded_train = pad_sequences(sequences_train, padding='post', maxlen=50)
```

Fig. 15 Numeric Sequence Conversion cell.

Texts are converted into numeric sequences, and post padding (padding='post') is applied to normalize the length to 50 tokens per message. The same procedure is applied to the validation data.

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Embedding, Bidirectional, LSTM, Dense

model = Sequential([
    Embedding(input_dim=10000, output_dim=64, input_length=90),
    Bidirectional(LSTM(64)),
    Dense(1, activation='sigmoid')
])

model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

Fig. 16 Bi-LSTM Model Definition cell.

The model consists of:

- An Embedding layer to represent words.
- A Bidirectional LSTM layer with 64 units to capture patterns in both directions of the text.
- A final Dense layer with sigmoid activation for binary classification.
- The model is then compiled using binary_crossentropy as the loss function and adam as the optimizer.

```
history = model.fit(padded_train, labels_train, epochs=5, batch_size=32, validation_data=(padded_test, labels_test))
```

Fig. 17 Bi-LSTM Model Training cell

The model is trained for 5 epochs using batches of 32 samples.

IV. EXPERIMENTS

On the other hand, the Model Validation process consists of validating the Bi-LSTM model, also carried out in Google Colab, using the 3,000 messages that were not used in the Model Learning process. That is, an interactive interface was coded within the Google Colab notebook, aligned with the Bi-LSTM model trained during Model Learning, so that a user can input any phrase and obtain as output whether the sentence is violent or not.

```
# Ingresar texto manualmente
texto_usuario = input("Ingresa un mensaje para analizar: ")

# Preprocesar el texto
secuencia = tokenizer.texts_to_sequences([texto_usuario])
padded_seq = pad_sequences(secuencia, padding='post', maxlen=18)

# Predecir
prediccion = model.predict(padded_seq)

# Mostrar resultado
resultado = int(prediccion[0][0] > 0.5)
print(f"\n Resultado del modelo: {'VIOLENTO (1)' if resultado == 1 else 'NO VIOLENTO (0)'}")
```

```
Ingresa un mensaje para analizar: Necesito q pares. Este ambiente es insoportable.
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 52ms/step

 Resultado del modelo: NO VIOLENTO (0)
```

Fig. 18 Interactive interface of the Bi-LSTM model in Colab.

Dataset Class Distribution

The dataset was intentionally constructed with balanced class proportions to avoid bias: the training set contained 9,000 violent and 9,000 non-violent messages, while the validation set comprised 1,500 violent and 1,500 non-violent messages, yielding an exact 50/50 distribution across both splits.

Evaluation with Classification Report

```
5/5 ━━━━━━━━━━━━━━━━━━━━ 1s 93ms/step
              precision    recall  f1-score   support

           0       0.90      0.91      0.89      1300
           1       0.91      0.91      0.89      1700

    accuracy                           0.90      3000
   macro avg       0.91      0.91      0.89      3000
weighted avg       0.91      0.91      0.89      3000
```

Fig. 19 Classification report of the Bi-LSTM model on the validation set.

The model achieved strong performance on the validation set, with 90% precision for non-violent messages and 91% precision for violent messages. It also achieved 91% recall for both classes, demonstrating a high capacity to correctly detect true cases. The F1-score was 0.89 in both categories, reflecting a good balance between precision and recall. Overall, the model's accuracy was 90%, indicating that 9 out of 10 predictions were correct.

Confusion Matrix of the Bi-LSTM model

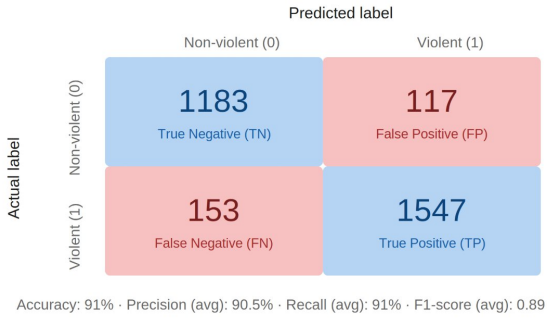

Fig. 20 Confusion matrix of the Bi-LSTM model on the validation set (n = 3,000).

Figure 20 presents the confusion matrix obtained from the validation set, the model correctly classified 1,183 non-violent messages and 1,547 violent messages, with 117 false positives and 153 false negatives, confirming a balanced performance across both classes.

Training progress curves of the Bi-LSTM

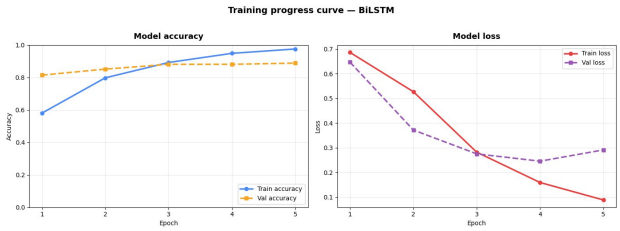

Fig. 21 Training progress curves of the Bi-LSTM model over 5 epochs: accuracy (left) and loss (right) for training and validation sets.

Fig. 21 illustrates the training and validation curves across 5 epochs, confirming that the model converged without overfitting. The validation accuracy stabilized around 0.89–0.90 from epoch 3 onward, while the validation loss remained consistently lower than the training loss, indicating good generalization.

Performance comparison of LSTM, GRU, and Bi-LSTM models

To validate the superiority of the Bi-LSTM architecture, two additional baseline models were trained under identical conditions — same dataset, tokenizer, padding length, and number of epochs — using a standard LSTM and a GRU network.

Fig. 22 Performance comparison of LSTM, GRU, and Bi-LSTM models across accuracy, precision, recall, and F1-score metrics.

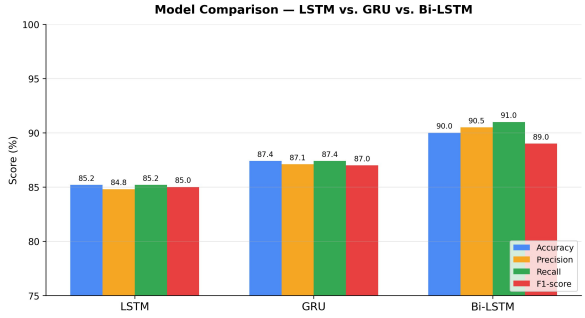

Fig. 23 Performance comparison of LSTM, GRU, and Bi-LSTM models on the validation set (n = 3,000).

Interpretation of Semantic Context in Violence Detection

This section shows that the model can correctly interpret the context in which potentially violent words are used. Phrases such as “cuando te vea te juro que te mato” (“when I see you, I swear I’ll kill you”) or “te voy a matar porque eres una inútil” (“I’m going to kill you because you’re useless”) are classified as violent due to their explicitly threatening tone. In contrast, expressions like “te voy a matar a besos” or “te voy a matar pero a besos” (“I’m going to kill you with kisses”) are recognized as non-violent. This demonstrates that the model is not limited to identifying keywords; it analyzes their intent within the message, reducing errors from literal interpretation.

Word “matar” (to kill)

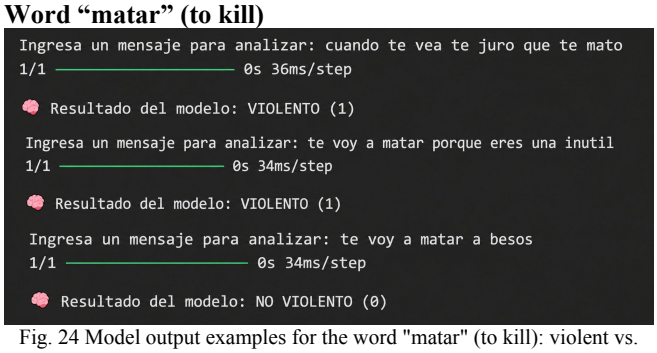

Fig. 24 Model output examples for the word "matar" (to kill): violent vs. non-violent interpretations.

Word “golpear” (to hit)

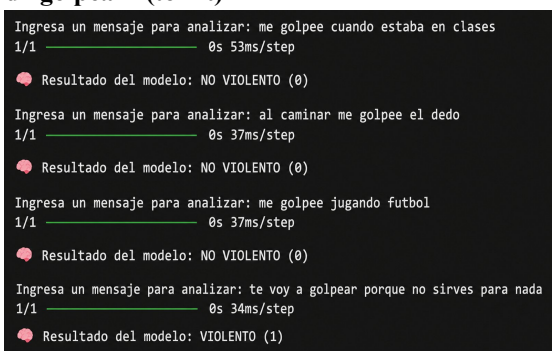

Fig. 25 Model output examples for the word "golpear" (to hit).

Word “dolor” (pain)

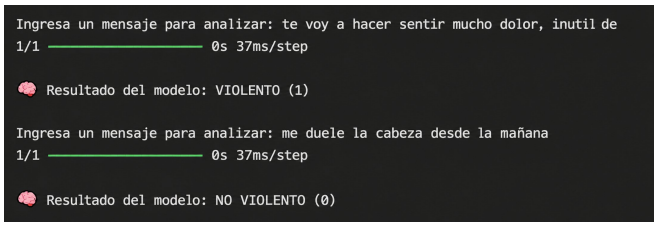

Fig. 26 Model output examples for the word "dolor" (pain): violent vs. non-violent interpretations.

Classification of Everyday Phrases: Identifying Neutral Content

This section shows that the model appropriately recognizes everyday, informational messages without violent content. Phrases such as “mañana llevo a mi gato al veterinario” (“tomorrow I’m taking my cat to the vet”), “hoy desayuné café con leche” (“today I had coffee with milk for breakfast”), or “¿quieres venir a almorzar con mi mamá?” (“do you want to come have lunch with my mom?”) are correctly classified as neutral. Even in the face of expressions with some indirect force—such as “el jefe necesita los informes ahora” (“the boss needs the reports now”) or “era demasiado trabajo para mí” (“it was too much work for me”)—the model does not infer aggression. This ability to avoid false positives is key to building a reliable and balanced system.

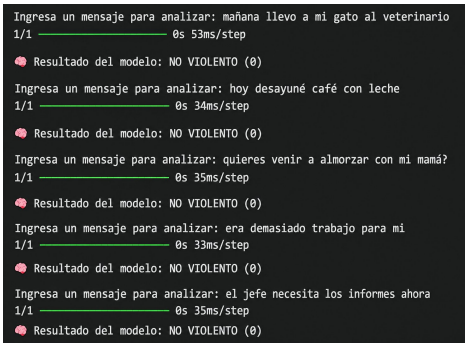

Fig. 27 Model output examples for everyday neutral phrases: correct non-violent classification.

Contextual Detection of Local Expressions in Peruvian Spanish

This section evaluates the model's performance with common expressions in Peruvian Spanish, both positive and negative. It correctly identifies colloquial phrases such as "jsjsjs," "eres lo máximo" ("you're the best"), or "jajaja qué abuso" (an ambiguous colloquialism, here treated as non-violent) as non-violent, even when there is cultural ambiguity. It likewise accurately detects negatively charged messages such as "siempre me decepcionas, inútil" ("you always disappoint me, useless") or "no sirves para nada, inútil" ("you're good for nothing, useless"). This shows that the model is not limited to standard structures, but recognizes patterns specific to Peruvian Spanish, distinguishing between neutral and offensive expressions—crucial for real-world, local use.

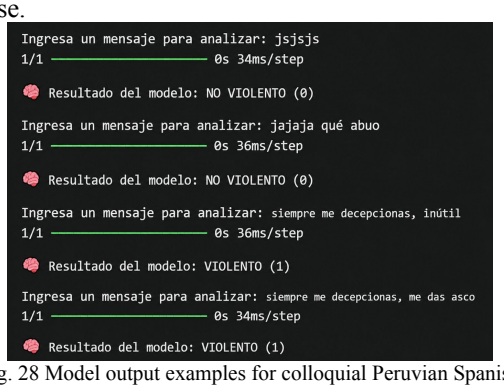

Fig. 28 Model output examples for colloquial Peruvian Spanish expressions: violent and non-violent classifications.

Ethical Considerations

This study was conducted in accordance with ethical principles applicable to the use of public digital data. The dataset was collected via web scraping exclusively from publicly available Twitter posts, without accessing any private or restricted content. All data were anonymized prior to processing, with personal identifiers such as usernames, locations, and URLs removed to ensure user privacy. The collected data were used solely for academic purposes, with no intention of identifying or profiling individuals.

Data collection and usage complied with the platform's Terms of Service and with established best practices in artificial intelligence research. No sensitive personal information was intentionally gathered. Dataset labeling was based on criteria defined by the authors according to linguistic patterns characteristic of the Latin American context, acknowledging the inherent subjectivity of this process. For ethical reasons, the dataset will not be publicly released in its original form.

This study did not involve direct experimentation with human subjects.

V. CONCLUSIONS AND EXPECTATIONS

Unlike traditional approaches that rely on generic datasets or models trained on neutral Spanish, our model is specifically designed to capture diverse linguistic and cultural particularities, with an initial focus on Peruvian Spanish while retaining the flexibility to adapt to any cultural and regional context. This adaptation is non-trivial, since expressions of violence and affection in local contexts are often laden with nuances that international models fail to interpret correctly. Colloquial phrases such as "I'll smash your face" (te parto la cara) or "I'm going to kill you with kisses" (te voy a matar a besos) require context-sensitive semantic analysis. Accordingly, the model we developed not only detects violence accurately but also reduces false positives by understanding the country's emotional and linguistic context. This local tailoring is a crucial contribution to automated risk detection in regions where verbal violence is normalized or veiled by everyday slang.

As part of our real-world, context-adapted approach, this study proposes doing away with reliance on surveys, interviews, or other conventional data-collection methods, which are often costly and limited in scope. Instead, we implement an automated pipeline based on web scraping of social platforms such as Twitter, accessing real messages from everyday digital environments. This strategy improves scalability and ensures spontaneous, representative language, demonstrating that it is possible to train robust models without intrusive or tightly controlled techniques.

The proposal further stands out for its ability to classify messages as violent or non-violent using textual content alone. This property is key for preventive applications where early risk detection can make a difference in cases of domestic violence or cyberbullying. Experimental results confirmed this capacity: the Bi-LSTM outperformed both LSTM (85.2% accuracy) and GRU (87.4% accuracy) baselines, achieving 90% accuracy and an F1-score of 0.89, with balanced performance across both classes.

Despite these results, the model exhibited limitations when processing sarcastic or highly context-dependent expressions common in Latin American informal speech. Phrases such as *'ay sí, eres lo máximo'* or *'qué inteligente que eres, de verdad'* — which carry mocking or threatening intent when used sarcastically — may be classified as non-violent due to their superficially positive wording. Similarly, expressions like

'sigue así nomás, ya verás' convey implicit threat through cultural subtext that the model cannot yet reliably capture. Incorporating sarcasm-aware training examples and contextual embeddings such as multilingual BERT represents a promising direction for future research.

Looking ahead, we expect future technological developments to use our model as the functional core of mobile applications or digital platforms that detect violent messages in real time. Given its context-sensitive accuracy in analyzing Latin American Spanish, it is well suited for integration into preventive tools that run in the background and issue alerts in response to aggressive patterns in school, family, or clinical settings.

In addition, we suggest that future research automate the entire pipeline of collection, cleaning, translation, and model retraining to ensure continual adaptation to the evolving digital lexicon. Such automation would help sustain performance without the need for ongoing manual tuning. We also recommend broadening data sources to include other social platforms, such as Facebook or Instagram, to enrich training with diverse communicative styles.

Finally, we anticipate that the developed model can be reused in international research thanks to its flexible architecture. This would allow adaptation to different languages and sociocultural contexts by adjusting corpora and key expressions to each setting, thereby facilitating automated analysis of verbal violence in regions beyond Peru while preserving the model's core architecture.

REFERENCES

[1]. Fernandez-Mamani, M., Janampa-Calderon, E., Conde, I., & Cjuno, J. (2024). Prevalence of domestic violence among health center users: a retrospective longitudinal analysis. Interacciones. https://doi.org/10.24016/2024.v10.428

[2]. Sánchez-Neyra, A., Lloclla-Gonzáles, H., & Silva-Díaz, H. (2024). Factors associated with suicidal ideation in adults from the Lambayeque Region, Peru, during the COVID-19 pandemic. Revista de la Facultad de Medicina Humana. https://doi.org/10.25176/rfmh.v24i1.6013

[3]. Melo, G., Bonafé, K., & Wachs-Lopes, G. (2022). Classification of depression using temporal text analysis in social network messages. Data Science and Machine Learning. https://doi.org/10.5121/csit.2022.121514

[4]. Ogunleye, B., & Dharmaraj, B. (2023). The use of a large language model for cyberbullying detection. Analytics, 2(3), 213–229. https://doi.org/10.3390/analytics2030038

[5]. Deas, N., Turcan, E., Pérez Mejía, I., & McKeown, K. (2024). MASIVE: Open-Ended Affective State Identification in English and Spanish. arXiv preprint. https://doi.org/10.48550/arXiv.2407.12196

[6]. Khan, S., Qasim, I., Khan, W., Khan, A., Khan, J., Qahmash, A., & Ghadi, Y. (2024). An automated approach to identify sarcasm in low-resource language. PLOS ONE, 19(4), e0307186. https://doi.org/10.1371/journal.pone.0307186

[7]. Tan, Y, Chow, C., Kanesan, J. et al. Análisis de sentimientos y detección de sarcasmo mediante aprendizaje profundo multitarea. Wireless Pers Commun 129, 2213–2237 (2023). https://doi.org/10.1007/s11277-023-10235-4

[8]. Ismail, W. (2024). Emotion Detection in Text: Advances in Sentiment Analysis Using Deep Learning. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 15(1), 17–26. https://doi.org/10.58346/JOWUA.2024.I1.002

[9]. Suguna, M., & Prabha, K. (2024). Reciprocating Encoder Portrayal from Reliable Transformer Dependent Bidirectional Long Short-Term Memory for Question and Answering Text Classification. IEEE Access, 12, 117800–117811. https://doi.org/10.1109/ACCESS.2024.3426604

[10]. Mahadevkar, S., Patil, S., & Kotecha, K. (2024). Enhancement of handwritten text recognition using AI-based hybrid approach. MethodsX, 12. https://doi.org/10.1016/j.mex.2024.102654

[11]. Wang, Y. (2024). Construction and improvement of English vocabulary learning model integrating spiking neural network and convolutional long short-term memory algorithm. PLoS ONE, 19(3 March). https://doi.org/10.1371/journal.pone.0299425

[12]. Yafooz, W. (2024). Enhancing Arabic Dialect Detection on Social Media: A Hybrid Model with an Attention Mechanism. Information (Switzerland), 15(6). https://doi.org/10.3390/info15060316

[13]. Zhao, Y., & Chen, D. (2021). Expression EEG multimodal emotion recognition method based on the bidirectional LSTM and attention mechanism. *Computational and Mathematical Methods in Medicine*, 2021. https://doi.org/10.1155/2021/9967592

[14]. Machová, K., Szabóová, M., Paralič, J., & Mičko, J. (2023). Detection of emotion by text analysis using machine learning. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1190326

[15]. Shanmugavadivel, K., Sampath, S., Nandhakumar, P., Mahalingam, P., Subramanian, M., Kumaresan, P., & Priyadharshini, R. (2022). An analysis of machine learning models for sentiment analysis of Tamil code-mixed data. *Computer Speech & Language, 76*, 101407. https://doi.org/10.1016/j.csl.2022.101407

[16]. Alomari, D., & Ahmad, I. (2024). Exploring character trigrams for robust Arabic text classification: A comparative analysis in the face of vocabulary expansion and misspelled words. *IEEE Access, 12*, 57103–57116. https://doi.org/10.1109/ACCESS.2024.3390048

[17]. Tahir, B., & Mehmood, M. (2021). Corpulyzer: A novel framework for building low resource language corpora. *IEEE Access, 9*, 8546–8563. https://doi.org/10.1109/ACCESS.2021.3049793

[18]. Luscombe, A., Dick, K., & Walby, K. (2021). Algorithmic thinking in the public interest: Navigating technical, legal, and ethical hurdles to web scraping in the social sciences. Quality & Quantity, 56(2), 1023–1044. https://doi.org/10.1007/s11135-021-01164-0

[19]. Cocón, F., Pérez-Cruz, D., Pérez-Rejón, J. Á., Zavaleta-Carrillo, P., Barradas-Arenas, U., Gómez-Ramón, R., & Pérez, J. (2023). *Web scraping: Uso de plataformas de extracción de datos aplicadas a un sitio web sobre profesiones en México*. RISTI: Revista Ibérica de Sistemas e Tecnologias de Informação, (52), 61–73. https://dialnet.unirioja.es/servlet/articulo?codigo=9694324

[20]. Gordon, W., & Rudin, R. (2022). Why APIs? Anticipated value, barriers, and opportunities for standards-based application programming interfaces in healthcare: Perspectives of US thought leaders. *JAMIA Open, 5*(1), ooac023. https://doi.org/10.1093/jamiaopen/ooac023

[21]. Pavlatos, C., Makris, E., Fotis, G., Vita, V., & Mladenov, V. (2023). Enhancing electrical load prediction using a Bidirectional LSTM neural network. Electronics, 12(22), 4652. https://doi.org/10.3390/electronics12224652