Natural Language Processing (2024), 1–20
doi:10.1017/nlp.2024.25

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Should we stay silent on violence? An ensemble approach to detect violent incidents in Spanish social media texts

Deepawali Sharma[1], Vedika Gupta[2], Vivek Kumar Singh[3,4] and David Pinto[5]

[1]Department of computer science, Banaras Hindu University, Varanasi, India, [2]Jindal Global Business Global School, O.P. Jindal Global University, Sonipat, Haryana, India, [3]Department of computer science, Banaras Hindu University, Varanasi, India, [4]Department of computer science, University of Delhi, Delhi, India, and [5]Faculty of Computer Science, Benemérita Universidad Autónoma de Puebla (BUAP). Mexico
Corresponding author: Vivek Kumar Singh; Email: vivek@bhu.ac.in

(Received 24 December 2022; revised 3 August 2023; accepted 23 November 2023)

Special Issue on '**Natural Language Processing Applications for Low-Resource Languages**'

Abstract

There has been a steep rise in user-generated content on the Web and social media platforms during the last few years. While the ease of content creation allows anyone to create content, at the same time it is difficult to monitor and control the spread of detrimental content. Recent research in natural language processing and machine learning has shown some hope for the purpose. Approaches and methods are now being developed for the automatic flagging of problematic textual content, namely hate speech, cyberbullying, or fake news, though mostly for English language texts. This paper presents an algorithmic approach based on deep learning models for the detection of violent incidents from tweets in the Spanish language (binary classification) and categorizes them further into five classes – accident, homicide, theft, kidnapping, and none (multi-label classification). The performance is evaluated on the recently shared benchmark dataset, and it is found that the proposed approach outperforms the various deep learning models, with a weighted average precision, recall, and F1-score of 0.82, 0.81, and 0.80, respectively, for the binary classification. Similarly, for the multi-label classification, the proposed model reports weighted average precision, recall, and F1-score of 0.54, 0.79, and 0.64, respectively, which is also superior to the existing results reported in the literature. The study, thus, presents meaningful contribution to detection of violent incidents in Spanish language social media posts.

Keywords: Deep learning; social media text analytics; Spanish language; violent incident detection

1. Introduction

There has been a steep rise in user-generated content on the Web and social media platforms during the last few years. The ease of content creation has resulted in user-generated content being created at an unprecedented rate. For example, the number of tweets created per second in 2022 was around 6,000.[a] Twitter is one of the most popular social network platforms, which allows people across the world to share different kinds of content. In fact, Twitter users might sometimes report real-time events quicker than news media. For instance, there have been existing

[a]https://www.dsayce.com/social-media/tweets-day/

studies to use Twitter as a tool to predict earthquakes sooner in neighboring regions.[b] It might also be used to alert about other natural and man-made disasters. However, at the same time, the volume and velocity of content generation make it difficult to monitor and control the spread of detrimental content. There have been instances of people using social media platforms to spread hate, post abusive or violent content, bully other users, etc. Therefore, it is very important to design algorithmic approaches for the automated identification of such content. The focus of the current study is violent incident detection.

Violent incidents include acts of terrorism, assaults, or any act that cause any physical injury to anyone. Violence has bad effects on those who witness or experience violence (Arellano *et al.* (2022)). Social media can play a crucial role in the early detection of violent events. Different law enforcement agencies can tap into social media channels and monitor posts that may describe violent incidents. This may help them in acting promptly to maintain the law and order. However, the large volume of social media content makes it very difficult to maintain a manual oversight of the social media channels and that is why there is a need for automated methods for the purpose. Several studies have been carried out to detect the specific violence in tweets and social media comments in the English language; for instance, sexual violence detection on Twitter using the #MeToo hashtag (Khatua *et al.* 2018; Hegarty and Tarzia 2019; Lopez *et al.* 2019). Similarly, some studies focused on detecting domestic violence and intimate partner violence from social media comments and posts (Subramani *et al.* 2017; Chen *et al.* 2020; Annapragada *et al.* 2021). There are also few studies on violence detection in content from languages other than English, such as in Arabic language (Abdelfatah *et al.* 2017; ALSaif and Alotaibi 2019; Khalafat *et al.* 2021). Detection of violent incidents and classifying the tweets into different categories in social media posts, particularly in non-English languages, is not that well explored, as we will see in the next paragraph.

Spanish belongs to the Indo-European language family. There exist more than 500 million native speakers of the Spanish language[c] globally, though it is mainly spoken in the United States of America, Spain, Mexico, etc. Spanish is a scarce or low-resource language. Numerous studies have been carried out in the Spanish language, particularly on hate speech detection (Basile *et al.* 2019; i Orts 2019; García-Díaz *et al.* 2023), cyberbullying (León-Paredes *et al.* 2019; Mercado *et al.* 2018; Cumba-Armijos *et al.* 2022), and offensive comments detection (Díaz-Torres et al. 2020; Ranasinghe and Zampieri 2021; Arango *et al.* 2022). However, there are very few studies on violent incident detection, more so in Spanish social media texts. Motivated by the research gap, this paper attempts to propose an algorithmic framework for the detection of violent incidents reported in social media texts in Spanish. First, the violent and nonviolent tweets are identified, followed by classifying the violent tweets into five categories, namely accident, homicide, theft, kidnapping, and none.

A set of deep learning models (convolutional neural network (CNN), long short-term memory (LSTM)) using GloVe embedding, transformer-based models (mBERT, XLM-RoBERTa, and BETO), and also an ensemble model (by combining CNN-BiLSTM, mBERT, XLM-RoBERTa, and BETO) are explored for the task of violent incident detection and subsequent categorization. More specifically, the following tasks are performed:

- First, some deep learning models (CNN and LSTM using GloVe embedding) and transformer-based models (mBERT and XLM-RoBERTa) for both the subtasks (refer to sections 4.1 and 4.2) are applied. The effect of using a language-specific model (BETO) for detecting violent incidents is also explored (Refer to section 4.2.3).

[b] https://digital.gov/2015/06/26/tweets-earthquakes/
[c] https://www.statista.com/statistics/991020/number-native-spanish-speakers-country-worldwide

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

- Second, an ensemble model (combining CNN-BiLSTM, mBERT, XLM-RoBERTa, and BETO) is designed and implemented for the two subtasks (refer to section 4.3).
- Finally, the results of the various models are compared with each other and also with the reported results of several previous studies (refer to section 5).

The rest of the paper is organized as follows: Section 2 discusses the related work for violent incident identification, particularly in the Spanish language. Section 3 describes the dataset, the tasks involved, and preprocessing steps. Section 4 presents the implementation of different methods and the final adopted methodology. Section 5 reports the results achieved, evaluation, and comparison. The paper concludes in Section 6 with a summary of the results and some possibilities of future work in the area.

2. Related work

Natural language processing research in the Spanish language has attracted the attention of researchers mainly during the last few years. Some studies focused on the detection of hate speech in Spanish as well as other low-resource languages (Basile et al. 2019; i Orts 2019; García-Díaz *et al.* 2023; Sharma *et al.* 2024b), cyberbullying (León-Paredes *et al.* 2019; Mercado *et al.* 2018; Cumba-Armijos *et al.* 2022), offensive comments (Díaz-Torres *et al.* 2020; Ranasinghe and Zampieri (2021); Arango *et al.* (2022)), and abusive content (Nobata *et al.* 2016; Prasanth *et al.* 2022; Sharma *et al.* 2024a). However, there are relatively very few studies on violent incident detection in Spanish. Nonavailability of a benchmark dataset could have been one of the main reasons. The first major dataset in the Spanish language for violent incident detection was provided by Iberian Language Evaluation Forum (IberLEF2022).[d]

Deep learning-based methods have been a popular choice recently for the automatic detection of events or features in social media texts. One recent study (Ta *et al.* 2022b) worked toward a classification of violent events in tweets in the Spanish language by using deep learning techniques. This study applied Multi-Task Deep Neural Network (MT-DNN) and used some preprocessing to remove or normalize the unwanted components and make the distribution of categories more balanced. The obtained F1-score, precision, and recall for subtask 1 are 74.80%, 75.52%, and 74.09%, respectively. Similarly, the F1-score, precision, and recall for subtask 2 are 39.20%, 37.79%, and 43.38%, respectively. Another study (Qin *et al.* 2022) used the multi-label classification framework based on prompt learning. This study used BiLSTM and CNN for multi-label classification and reported macro-F1 = 0.554281, recall = 0.562260, and precision = 0.550030.

For the past few years, transformer-based models have been preferred over classical machine learning and conventional deep learning approaches. Considering the Spanish language, some studies experimented with different flavors of transformer-based models. One of them applied distilled version of the BETO (DistilBETO) pretrained model (Tonja *et al.* 2022). The study demonstrated that the language-specific model (BETO) can outperform multi-lingual models. Similarly, another study used the transformer-based model (GAN-BERT) based on the BERT architecture used for the first subtask (Ta *et al.* 2022a). Back translation and preprocessing have been done to handle the tweets in the Spanish language. The model was tried on four runs with different hyperparameters and obtained an F1-score of 0.74, precision of 74.08%, and recall of 74.79%. Another study (Turón *et al.* 2022) translated the Spanish tweets into another language and passed them to a multilingual model. It experimented with mBERT, BETO, and Massive Artificial Intelligence (MarIA) models to classify the Spanish language tweets.

[d] https://codalab.lisn.upsaclay.fr/competitions/2638#participate-get-data

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

Table 1. Tabular representation of previous research work

		Results					
		Subtask 1			Subtask 2		
Author	Approach	Precision	Recall	F1-score	Precision	Recall	F1-score
Vallejo-Aldana *et al*. (2022)	Ensemble model (multi-tasking learning approach)	0.8032	0.7503	0.7759	0.47	0.47	0.4733
Tonja *et al.* (2022)	Transformer-based model (DistilBETO)	0.76	0.73	0.7455	0.49	0.49	0.4903
Ta *et al.* (2022a, 2022b)	Transformer-based model (GAN-BERT)	74.08%	74.79%	74.43%	–	–	–
Montanés-Salas *et al.* (2022)	Voting ensemble approach	–	–	0.76	–	–	0.50
Qin *et al.* (2022)	Prompt-based framework (Prompt + BiLSTM + CNN + Assyloss)	–	–	–	0.550030	0.562260	Macro F1 = 0.554281
Ta *et al.* (2022a, 2022b)	Multi-Task Deep Neural Network (MT-DNN)	75.52%	74.09%	74.80%	37.79%	43.88%	39.20%
Turón *et al.* (2022)	Transformer-based model (BETO, mBERT, MarIA)	–	–	77.32%	–	–	52.86%

Some recent previous studies have also used ensemble approach to detect violent incidents from tweets. One such study (Vallejo-Aldana *et al.* 2022) used the multi-task learning approach to ensemble the three different models with the different secondary tasks for each. The secondary task chosen for the individual models was done on their individual performance. Similarly, another study (Montañes-Salas *et al.*, 2022) experimented with different transformer-based models (Optimized BETO, Optimized BSC Roberta, Optimized Twitter XLM, and Optimized BETO + Preprocessing) and implemented a standard majority voting for both subtasks.

A summarized view of the relevant previous studies is presented in a tabular form in Table 1. As can be seen, there is still a need for improving the accuracy and model performance further. This work, therefore, presents an attempt in this direction.

3. Dataset description and preprocessing

The dataset used in this paper is provided by Iberian Language Evaluation Forum (IberLEF2022) called DA-VINCIS . This dataset contains tweets written in the Spanish language. Figure 1 (a) and (b) illustrate the two subtasks evaluated in this paper. Both subtasks are performed on the same dataset. The dataset contains two attributes: tweets and incident detection. Here, the tweet attribute contains the tweets from the user that are extracted from Twitter and the incident detection attribute contains the category of the tweets whether the tweet is a violent incident (1) or not (0). The first subtask is binary classification to identify whether a tweet is a violent incident or not.

If the tweet belongs to the violent category, then it is represented by 1, otherwise 0. The dataset contains 3,362 tweets of which 1,798 belong to the nonviolent incident and 1,564 belong to the violent incident category. Table 2 shows the distribution of tweets for both categories: 0 and 1.

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

Table 2. Distribution of tweets for each category

Category	Number of tweets
0	1798
1	1564
Total	3362

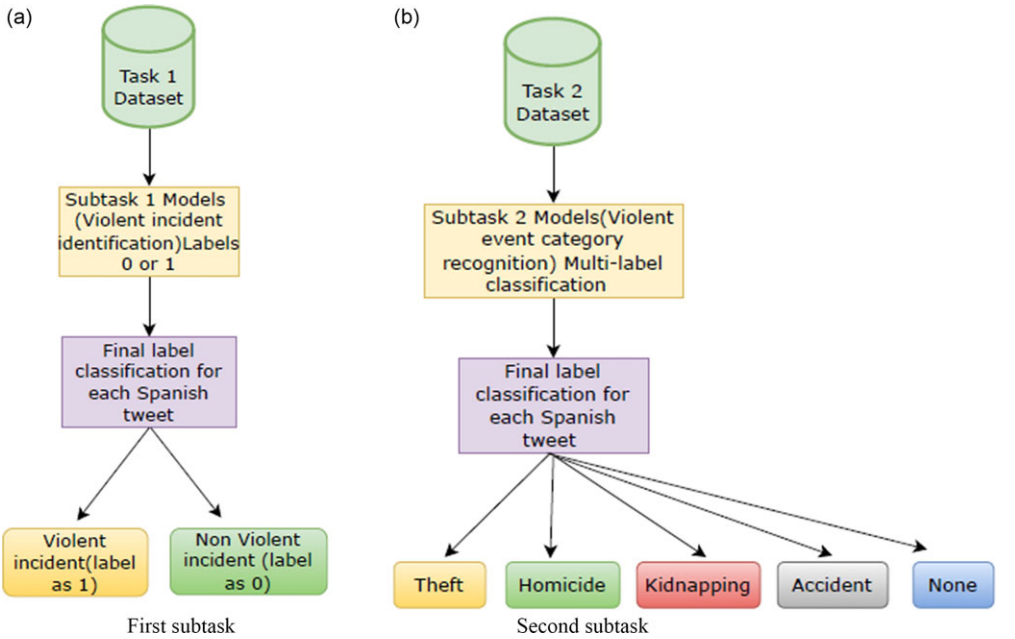

Figure 1. Illustration of the subtasks performed.

Two of the randomly selected tweets from the dataset (one for each category) and their translation in English are shown below as an illustration:

"Casos como el accidente en Villa Lorena, en el que por un lado tenemos al responsable de las muertes por su imprudencia, pidiendo clemencia; y por el otro las victimas pidiendo un resarcimiento a su dolor, evidencian la necesidad de establecer procesos de justicia restaurativa". – 0 (Nonviolent incident)

English translation: Cases like the accident in Villa Lorena, in which on the one hand we have the person responsible for the deaths due to his recklessness, asking for mercy; and on the other hand, the victims asking for compensation for their pain show the need to establish restorative justice processes.

"Dan el último adiós en Gaira a Laura De Lima y Juan Alzate, víctimas del accidente en Santa Marta". – 1 (Violent incident)

English translation: They give the last goodbye in Gaira to Laura De Lima and Juan Alzate, victims of the accident in Santa Marta.

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

Table 3. Category-wise distribution of tweets

Category	Number of tweets
Accident	1125
Homicide	260
Kidnapping	45
Theft	179
None	1798

Table 4. Category-wise distribution of tweets

Category	Original Tweets	Translation
Accident	*"#Chiriqui¬ — Atendemos accidente vehicular, vía hacia Santo Tomas, Bugaba. Motorizado y pasajero chocan contra objeto fijo, trasladados a centro médico por paramédicos del #BCBRP https://t.co/opy7pkmLqJ"*	#Chiriqui — We attend vehicular accident, road to Santo Tomas, Bugaba. Motorized and passenger crash against a fixed object, transferred to a medical center by #BCBRP paramedics https://t.co/opy7pkmLqJ
Homicide	*"Este domingo fueron localizados los cuerpos de dos hombres en Caminos a San Isidro Mazatepec y Juárez en el Poblado de La Cofradía de la Luz, presentaban huellas de violencia, el @forensesjalisco realiza el levantamiento correspondiente @zona3noticias @OPEALERTCR @REDTNJalisco https://t.co/Rmb8TqUSB6"*	This Sunday the bodies of two men were located on Caminos a San Isidro Mazatepec and Juárez in the town of La Cofradía de la Luz, they showed signs of violence, the @forensesjalisco carried out the corresponding survey @zona3noticias @OPEALERTCR @REDTNJalisco https://t.co/Rmb8TqUSB6
Kidnapping	*"Capturan a defensores de derechos humanos que habrían participado en secuestro! Los defensores de derechos humanos detenidos tenían protección del Estado y se movilizaban en dos camionetas que fueron quemadas. https://t.co/6Vm7ZsHpLU"*	They capture human rights defenders who would have participated in a kidnapping! The detained human rights defenders had state protection and were traveling in two vans that were burned. https://t.co/6Vm7ZsHpLU.
Theft	*"# ÚltimaHora Llega al CERESO "la gaviota", por robar y desatar balacera en el Morelos; la acusan de robo calificado y lesiones dolosas.@JLMNoticias #LuceroÃ lvarez https://t.co/oQ7zMgvTR2"*	# ÚltimaHora "la gaviota" arrives at CERESO, for stealing and unleashing shootings in Morelos; They accuse her of qualified robbery and intentional injuries.@JLMNoticias #LuceroÃlvarez https://t.co/oQ7zMgvTR2
None	*"Casos como el accidente en Villa Lorena, en el que por un lado tenemos al responsable de las muertes por su imprudencia, pidiendo clemencia; y por el otro las victimas pidiendo un resarcimiento a su dolor, evidencian la necesidad de establecer procesos de justicia restaurativa. https://t.co/vPTrxnhCqA"*	Cases like the accident in Villa Lorena, in which on the one hand we have the person responsible for the deaths due to his recklessness, asking for mercy; and on the other hand, the victims asking for compensation for their pain, show the need to establish restorative justice processes. https://t.co/vPTrxnhCqA

The second subtask is a multi-label classification that identifies what type of violent incident is reported by the tweet. These categories are accident, homicide, kidnapping, theft, and none. Table 3 shows the category-wise distribution of tweets.

Randomly selected tweets from each category and their translation in English are given in Table 4 for illustration.

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

3.1 Data preprocessing

Preprocessing is an important step of various natural language processing (NLP) tasks. In the present work, the following steps are followed to preprocess the tweets:

1. **Removal of punctuation marks**: All punctuation marks such as question marks, commas, or quotation marks are removed as they do not add any valuable information.
2. **Removal of stop words**: By using the nltk library, all those words that are not providing meaningful information are removed.
3. **Tokenizing the text**: The method tokenize(), tokenizes the tweets into tokens or words.
4. **Removing URLs**: Some tweets contain URLs to the website and are removed as they only make the text noisier.
5. **Removing hashtags**: All hashtags are removed from the tweets.
6. **Remove digits**: For violent identification, numbers are not providing any information and hence they are also removed.
7. **Remove repeated characters**: On social media, it is common to repeat the characters at the end of the word. So, it is needed to remove these extra characters.
8. **Lemmatization**: In this process, all the words are replaced with their dictionary form or root word. It is performed to enable the pipeline to treat the present and past tense by considering the context surrounding a word. This helps to decide which root is accurate when the word form alone is vague.
9. **Removing blank spaces**: The last step is removing extra blank spaces that could have been created when words are deleted in previous steps, or they may have already been there in the tweet.

4. Methodology

This section presents the overall methodology followed in this work along with details of the algorithmic models. The most prevalent choice of technique is to use deep learning-based models like CNN and Bidirectional Long-Short-Term Memory (Bi-LSTM). Both can be used individually to detect and classify violence in Spanish language posts. These implementations are described briefly. Next, the implementation of the currently popular approach of transformer-based models is described. This approach has been proven to be better than other models for many such tasks. The transformer-based models are applied individually, namely BERT, BETO, and XLM-RoBERTa. Finally, the ensemble model proposed for the task by combining several models is presented. For the purpose of developing the proposed ensemble model, Bi-LSTM is used in combination with CNN. The motivation behind taking this combination is that the CNN extracts important features from text through pooling but misses fully extracting contextual information in both directions. This is where the role of Bi-LSTM comes into play. Bi-LSTM helps to fetch the information sequentially in both directions (backward or forward). Therefore, the combined setup of CNN and Bi-LSTM gives improved results as compared to the stand-alone single model.

4.1 Deep learning-based models

4.1.1 CNN

CNN is a type of neural network used for image or object detection and classification. For text classification, a one-dimensional convolutional layer and a word embedding layer are required. Word embedding represents the word as a vector. It maps semantically similar words. The length of words in each sequence is different; thus, in order to specify the length of the word sequence, the parameter maxlen is used and pad_sequence() pads the sequence of words with zeroes. Figure 2

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

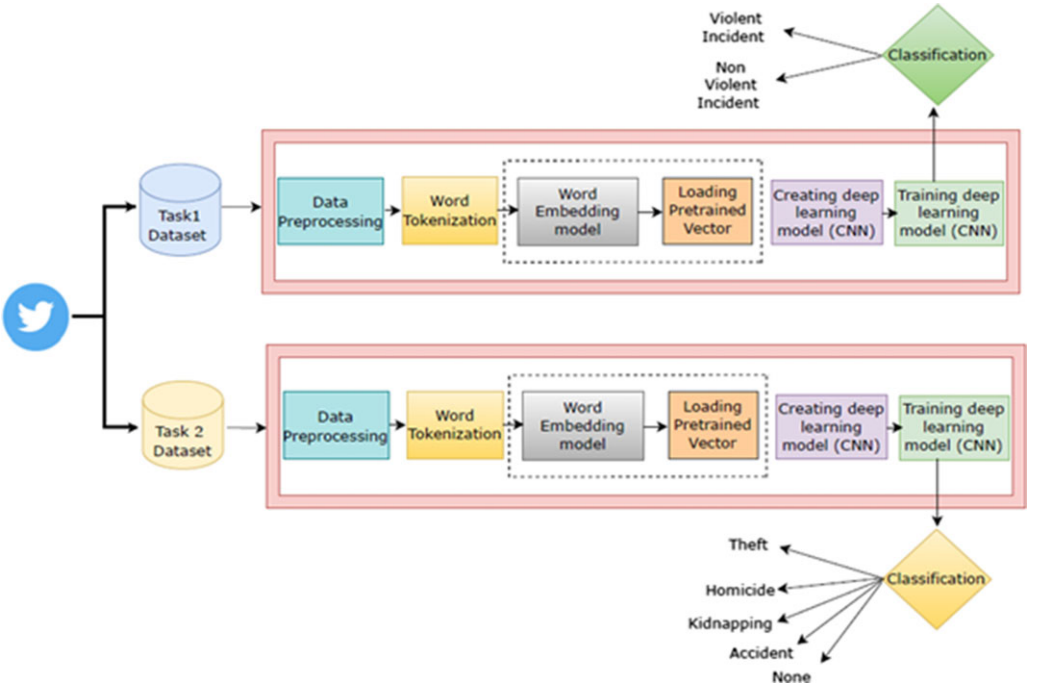

Figure 2. Implementation of CNN for both tasks.

shows the implementation of CNN for both datasets. In this paper, GloVe embedding is used. The embedding features are passed to the convolutional layer in which the ReLu activation function is used. After that, max-pooling layer is used to reduce the dimension complexity. The dropout technique is used with a dropout rate of 0.2 to remove the chance of overfitting and the final layer is a dense layer in which sigmoid is used as an activation function for the first subtask to classify the tweets as a violent incident or not (binary classification) and for the second subtask, softmax is used to classify the tweets in given five classes: Accident, Homicide, Kidnapping, Theft, and None (multi-label classification). The model is trained on 100 epochs using the "Adam" optimizer and binary cross-entropy loss for binary classification problems and categorical cross-entropy for multi-label classification.

4.1.2 LSTM

LSTM is a type of recurrent neural network (RNN) that can have multiple hidden layers and information passes through every layer. The relevant information is kept, and irrelevant information gets discarded in every single cell (Yu *et al.* 2019). After preprocessing the text, the tokens are fed to the embedding layer as shown in Figure 3. Each token is passed sequentially to the embedding layer in which words are represented as a dense vector. The next layer is the LSTM layer with 128 neurons, and the activation function is Rectified Linear Unit (ReLU). The return_sequences=TRUE is an important parameter to get the input for the next layer from the output of the previous layer. The dropout rate here is 0.2 to reduce the biasness. The last and final layer is a dense layer. The model is trained on 100 epochs using the "Adam" as an optimizer, and the batch size is 64. For the first subtask, the "binary cross-entropy" loss function is used and sigmoid is used as activation function to classify tweets as violent incidents or not. Similarly, for the second subtask, the "categorical cross-entropy" loss function is used and softmax as an activation function to classify the tweets in the given five categories.

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

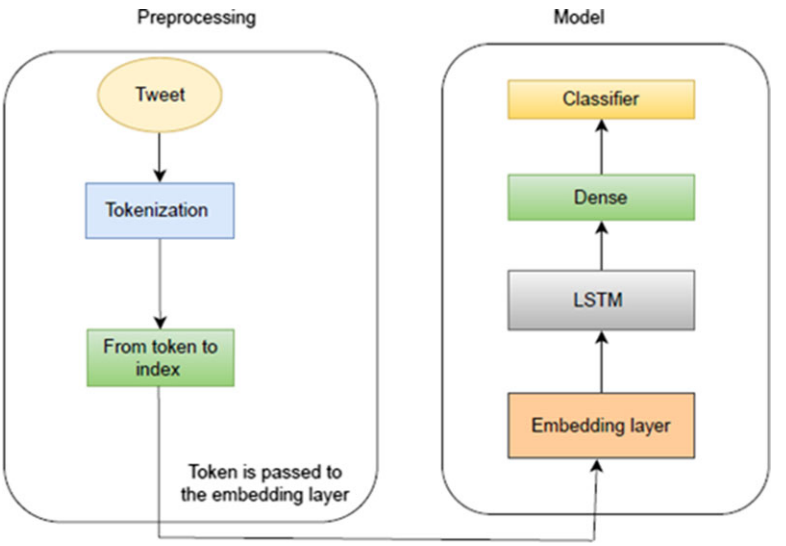

Figure 3. Block diagram to show the implementation of LSTM.

4.2 Transformer-based models

The transformer is an architecture to solve sequence-to-sequence tasks while handling long-range dependencies with ease (Vaswani *et al.* 2017). The transformer completely relies on self-attention to compute the input and output representation without using the sequence-aligned RNNs or convolution. Self-attention is also known as intra-attention. It is an attention mechanism to compute a representation of the sequence that relates different positions of a single sequence. Transformer architecture has two blocks – encoder and decoder as shown in Figure 4. The multiple identical encoders and decoders are stacked on top of one another. The number of units in each encoder and decoder stack is the same. Each encoder has two layers: multi-head attention and feed-forward neural network. Similarly, decoder has three layers: masked multi-head attention, multi-head attention, and feed-forward network. The input is represented by using embedding for both the encoder and decoder. The transformer injects a vector to individual input embeddings to preserve the positional information, and this vector is known as positional encoding. This injected vector is added to the embeddings at the bottom of both the encoder and decoder blocks. After that, the word embeddings are passed to the encoder and then transformed and passed to the next encoder and so on. Finally, the output that is obtained from the last encoder is passed to all the decoders.

4.2.1 mBERT

BERT stands for Bidirectional Encoder Representations from Transformers (Devlin *et al.* 2018). BERT is developed by Google and pretrained with the Wikipedia and Toronto Book corpus. BERT architecture consists of several layers of Encoder and each encoder consists of two sublayers: a feed-forward layer and a self-attention layer. First, the data are preprocessed (refer to section 3.1), and after preprocessing, the data are tokenized using the BERT tokenizer. Two special tokens are added in each sequence of tokens namely [CLS] and [SEP]. CLS is the first token of each sequence, and SEP is used to separate the segments as shown in Figure 5. The maximum length of sequence that can be fed to the BERT model is 512. If the length of the sequence is less than 512, then the unused slots are padded using the [PAD] token, and if the length of the sequence is more than 512 then the sequence is truncated.

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

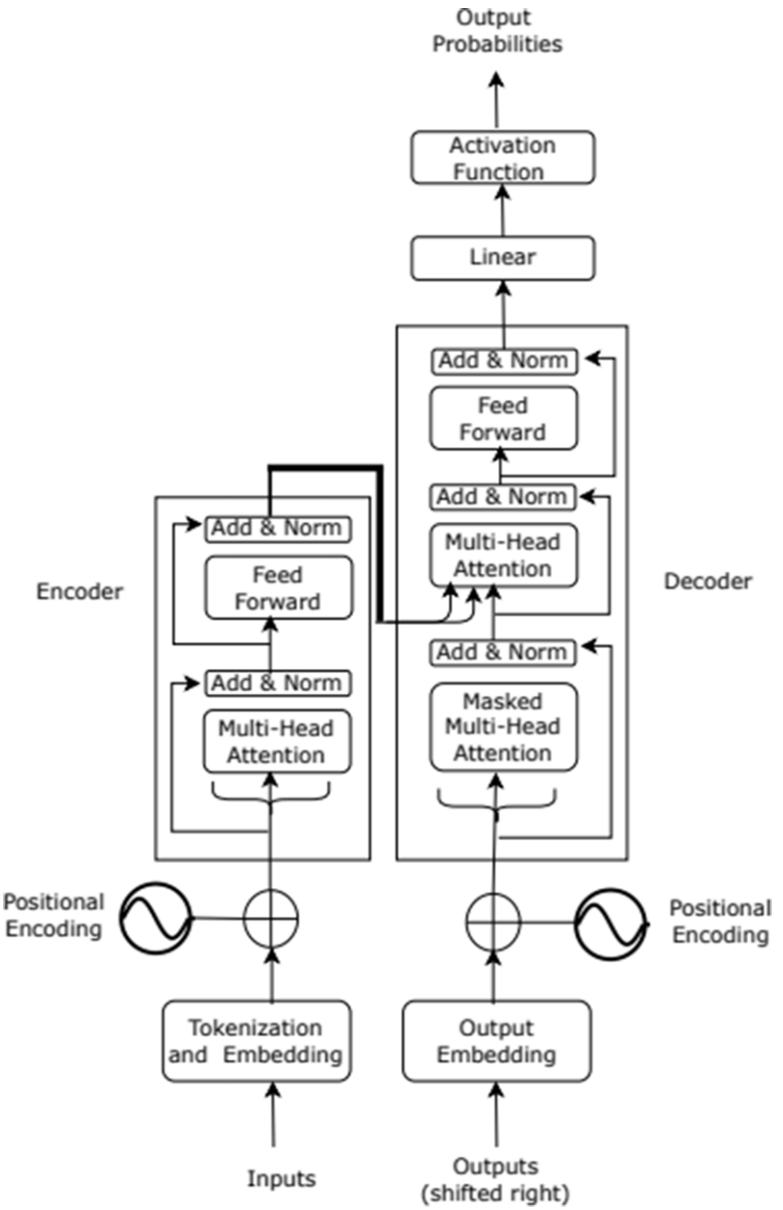

Figure 4. Architecture of Transformer.

The tokenized data is fed to the BERT model for classification. Figure 6 shows the block diagram of BERT. The model is trained on 10 epochs using the "AdamW" as an optimizer and the learning rate is set to be 3e-5. For the multi-label classification, cross-entropy loss is employed; and for the binary classification, binary cross-entropy with logits loss (BCEWithLogitsLoss) is used.

4.2.2 XLM-RoBERTa

XLM-RoBERTa is a multilingual language model based on Facebook's RoBERTa model (Conneau *et al.* 2019). RoBERTa is a transformer-based model, and its implementation is similar to BERT model (Liu *et al.* 2019). It is a transformer-based masked language model that is trained on 100

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

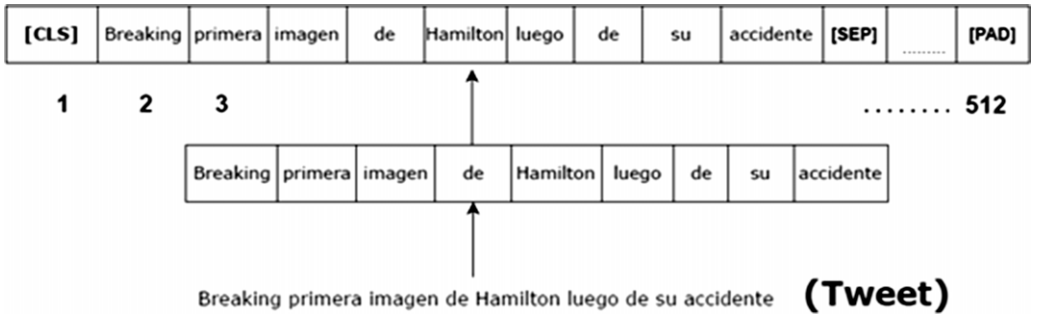

Figure 5. Illustration of BERT tokenizer.

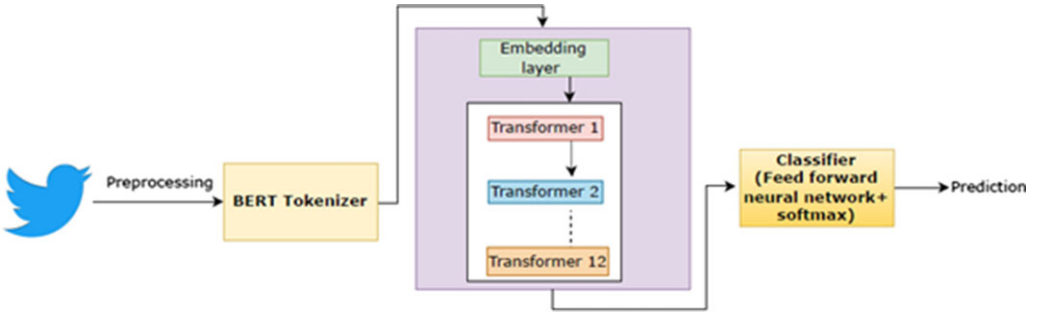

Figure 6. Block diagram of BERT.

different languages. RoBERTa is pretrained on a large corpus in a self-supervised manner. For training, more than two terabytes of filtered CommonCrawl data was used. The implementation of XLM-RoBERTa includes a preprocessing step, as described earlier. After preprocessing the data, the tweets are tokenized using the XLM-RoBERTa tokenizer. The token_ids are generated after tokenizing the tweets. The maximum length of the input text that BERT can be processed is 512 with special tokens: [CLS] and [SEP]. Now, the XLMRobertaForSequenceClassification function is called to use the pretrained xlm-roberta-base model. At last, the model was trained for 10 epochs with batch size 64 using the "AdamW" optimizer. The learning rate is set to 3e-5 in order to classify the tweets as violent incidents or not violent incidents (for the first subtask). The violent incidents are thereafter categorized into five classes (for the second subtask).

4.2.3 BETO

BETO is the Spanish version of BERT trained on big Spanish corpus. The size of BETO is like a BERT base, and it was trained with the Whole Word Masking technique.[e] Before starting the model training, the parameters of batch size and maximum length are set. The batch size is set to be 64, which is the number of samples used in one iteration of the training. The maximum sequence length is set to be 512. After that hyperparameters are adjusted to obtain good results. Hyperparameters are those parameters that control how much the model changes when the model weights are updated in response to the estimated error. Hyper indicates that these are "top-level" parameters that control the learning process. Hyperparameters that are adjusted are Learning rate, Epochs. The learning rate is set to be 3e-5, and the model is trained on 10 epochs to classify the tweets as violent incidents or not.

[e] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

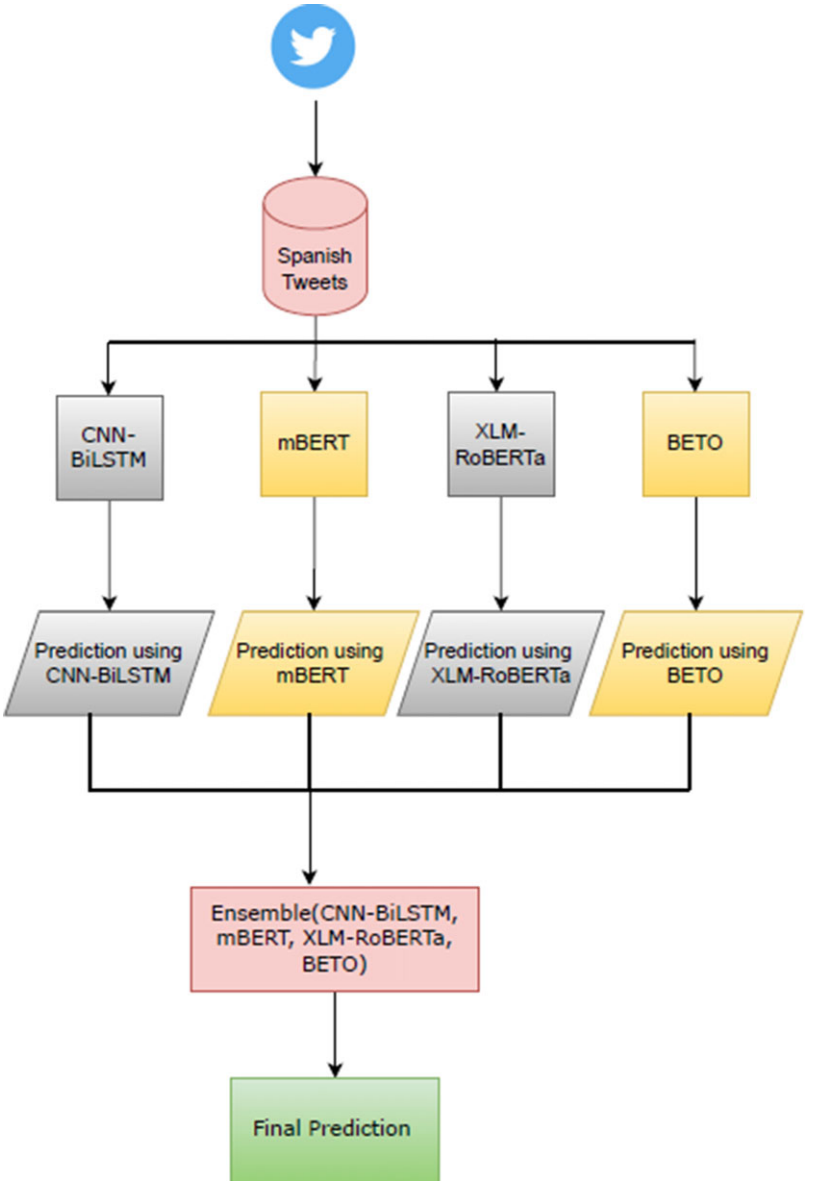

Figure 7. Ensemble approach for violent incident detection in Spanish

4.3 Proposed ensemble model

The study proposes an ensemble model by combining CNN-BiLSTM, mBERT, XLM-RoBERTa, and BETO models in an ensemble configuration. The idea is that through an ensemble, the advantages of the models can be combined. For developing the ensemble model, Bi-LSTM is combined with CNN to make a single model CNN-BiLSTM. The ensemble then combines the transformer-based models too in a simple voting scheme configuration. The configuration allows utilizing the performance of all the stand-alone models in a way that the best predictions come together to make the final prediction.

Figure 7 shows the complete ensemble approach for violent incident detection in the Spanish language tweets. The proposed model aggregates the prediction of deep learning- based

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

model (CNN-BiLSTM) and transformer-based models (mBERT, XLM-RoBERTa, and BETO) and ensemble them to get the final prediction. Each model is trained individually, and then the models are combined in an ensemble. When the ensemble approach is used, then the prediction error of the model reduces as long as the base models are varying and independent. Each model of the ensemble framework contributes equally by making a vote using their predictions. Voting Classifier is used to ensemble the models and select the predicted class using the simple majority voting scheme. The voting scheme works as follows: each of the individual models predict the class of a tweet as "0" or "1". All these votes for a given tweet are then seen together, and whichever class (out of "0" and "1") is preferred by the majority of the models is chosen as the final class for the given tweet. For example, let's assume that three models predict a tweet as violent incident (Label 1) and one model predicts the tweet as nonviolent incident (Label 0), then majority vote is for violent incident class and hence the tweet is assigned this label. However, since the ensemble has four models, there may be certain tweets for which votes for the two classes become equal. For example, for a given tweet, two models predict violent incident (Label 1) class and other two models predict nonviolent incident (Label 0) class. In this case of tie, the performance-based weighting is used, that is, higher weight is assigned to vote of a model having overall better performance.

The working of the model can be further illustrated with Algorithm 1.

Algorithm 1: Ensemble (CNN-BiLSTM, mBERT, XLM-RoBERTa, and BETO) model

(1) *Initialize an empty list to store predictions for each model: predictions_list = []*

(2) *Obtain prediction using each model:*

   (a) *Make a prediction using the CNN-Bi-LSTM model: cnn_bilstm_pred = CNN_BiLSTM.predict (tweet_data)*
   (b) *Make a prediction using the mBERT model: mbert_pred = mBERT.predict (tweet_data)*
   (c) *Make a prediction using the BETO model: beto_pred = BETO.predict (tweet_data)*
   (d) *Make a prediction using the XLM_RoBERTa model: xlm_roberta_pred = XLM_RoBERTa. predict (tweet_data)*

(3) *Initialize a list to store the final predictions: final_predictions = []*

(4) *Determine the final class of each tweet*

   (a) *Count the number of votes for each class.*
   (b) *If there is a majority vote for a class, add that class to the final_predictions list.*
   (c) *If there is a tie (equal number of votes for both classes), use performance-based weighting to determine the final class and add it to the final_predictions list.*

(5) *Return the final_predictions list as the output*

The above-mentioned Algorithm-1 can be more technically described in the following pseudocode, for the ease of reproducibility:

Pseudocode:

```
#Import all libraries
#Input data
#preprocess the data
def predict_cnn_bilstm():
    #prediction
    return prediction
def predict_mbert():
```

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

```
    #prediction
    return prediction
def predict_beto():
    #prediction
    return prediction
def predict_xlm_roberta():
    #prediction
    return prediction
def performance_based_weighting(predictions_list, weights):
    #weights to give higher importance to more reliable models
    #return the find weighted prediction
def ensemble_voting():
cnn_bilstm_model = predict_cnn_bilstm()
mbert_model = predict_mbert()
beto_model = predict_beto()
xlm_roberta_model=predict_xlm_roberta()
ensemble_predictions=VotingClassifier(estimators= [cnn_bilstm_model,mbert_model,beto_
model,xlm_roberta_model, Voting='hard')
weights=[] #performance based weights
final_prediction=performance_based_weighting (ensemble_predictions, weights)
return final_prediction
```

For the subtask 1 (binary classification), the model is trained for five epochs using the "AdamW" optimizer, "BCEWithLogitsLoss" as a loss function, and the learning rate set as 3e-5 to detect whether the incident is violent or not. Similarly, for subtask 2 (multi-label classification), the model is trained for five epochs using the "AdamW" optimizer, "CrossEntropyLoss" as loss function, and 3e-5 learning rate set to classify the violent incidents in given categories: accident, homicide, kidnapping, theft, and none.

5. Results

This section presents results and a comparative analysis of the performance of the deep learning-based and transformer-based models implemented for both the subtasks. The standard performance metric is computed. The results in the present study are also compared with previous relevant studies.

5.1 Experimental results

To report the results, three standard performance metrics – precision, recall, and F1-score are used. The metrics are defined as follows:

$$Precision(P) = TP/(TP + FP) \tag{1}$$

$$Recall(R) = TP/((TP + FN)) \tag{2}$$

$$F1 - score = (2 * P * R)/(P + R) \tag{3}$$

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

Table 5. Classification report for the implemented models on the first subtask

Model		CNN	LSTM	mBERT	XLM-RoBERTa	BETO	Proposed model (Ensemble of CNN Bi-LSTM, mBERT, BETO, XLM-RoBERTa)
Embedding		GloVe	GloVe	–	–	–	–
Violent incident	Precision	0.71	0.71	0.69	0.85	0.76	0.8
	Recall	0.64	0.68	0.78	0.59	0.7	0.86
	F1-score	0.67	0.7	0.73	0.7	0.73	0.83
Nonviolent incident	Precision	0.71	0.73	0.8	0.72	0.77	0.82
	Recall	0.77	0.76	0.71	0.91	0.82	0.75
	F1-score	0.74	0.75	0.75	0.8	0.79	0.76
Weighted Average	Precision	0.71	0.72	0.75	0.78	0.77	**0.82**
	Recall	0.71	0.72	0.74	0.76	0.77	**0.81**
	F1-score	0.71	0.72	0.74	0.75	0.77	**0.8**

where, TP, FP, and FN are True Positive, False Positive, and False Negative, respectively. The metric (precision, recall, and F1- score) are computed independently for each of the classes of both subtasks, and then the weighted average of them is reported.

The P, R, and F1-score values for the various methods implemented are presented for the first subtask. Table 5 shows the performance metrics of the two deep learning models (CNN and LSTM), transformer-based models (mBERT, XLM-RoBERTa, and BETO), and the Ensemble (CNN-BiLSTM, mBERT, XLM-RoBERTa, and BETO) for the first subtask.

For the first subtask, the five implemented models: CNN, LSTM, mBERT, XLM-RoBERTa, and BETO reported weighted average precision as 0.71, 0.72, 0.75, 0.78, and 0.77, respectively. The obtained weighted average recall reported is 0.71, 0.72, 0.74, 0.76, and 0.77 for CNN, LSTM, mBERT, XLM-RoBERTa, and BETO, respectively. The weighted average F1-score achieved for CNN, LSTM, mBERT, XLM-RoBERTa, and BETO are 0.71, 0.72, 0.74, 0.75, and 0.77, respectively.

On the other hand, the proposed Ensemble implementation achieves a weighted average precision, recall, and F1-score of 0.82, 0.81, and 0.80, respectively, which is better than the other implementations. Thus, for the first subtask, the proposed model outperforms the other models.

Table 6 shows the performance metrics for the various methods implemented on the second subtask. For the second subtask, the deep learning-based models (CNN and LSTM) reported weighted average precision as 0.59 and 0.61, respectively. The transformer-based models (mBERT, XLM-RoBERTa, and BETO) achieve weighted average F1- score as 0.63, 0.63, and 0.63, respectively.

In the case of the second subtask too, the proposed model outperforms the other models and achieves a weighted average precision, recall, and F1-score of 0.54, 0.79, and 0.64, respectively. The comparison of metric values for various implementations shows that the BETO model performs better than the mBERT as well as XLM-RoBERTa. In this context, the main difference between these models can be considered. BETO was trained with data written in Spanish, mBERT was trained with texts written in 104 languages, and XLM-RoBERTa was trained with data written in 100 languages. Among these models, the number of Spanish words used in the training differ greatly. This means that the percentage of words in the models' vocabulary is also very different. This may be the reason for a significant impact on the results.

It is also observed that the deep learning models performed poorer than transformer-based learning models possibly due to the following reason. In deep learning models, the pretrained

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

Table 6. Classification report for the implemented models on the second subtask

Model		CNN	LSTM	mBERT	XLM-RoBERTa	BETO	Proposed model
Embedding		GloVe	GloVe	–	–	–	–
Accident	Precision	0.39	0.46	0.5	0.51	0.51	0.53
	Recall	0.95	0.92	0.88	0.89	0.85	0.83
	F1-score	0.56	0.61	0.64	0.65	0.64	0.64
Homicide	Precision	0.12	0.14	0.16	0.23	0.17	0.18
	Recall	0.48	0.62	0.59	0.68	0.51	0.5
	F1-score	0.19	0.22	0.25	0.34	0.25	0.27
Kidnapping	Precision	0	0.33	0.08	0.75	0.07	0.1
	Recall	0	0.02	0.21	0.05	0.09	0.14
	F1-score	0	0.03	0.12	0.1	0.08	0.11
Theft	Precision	0.15	0.14	0.13	0.15	0.09	0.1
	Recall	0.12	0.34	0.39	0.4	0.17	0.19
	F1-score	0.14	0.2	0.19	0.22	0.12	0.13
None	Precision	0.55	0.57	0.6	0.57	0.63	0.65
	Recall	1	0.97	0.92	0.94	0.89	0.87
	F1-score	0.71	0.72	0.72	0.71	0.73	0.74
Weighted Average	Precision	0.44	0.48	0.5	0.51	0.52	**0.54**
	Recall	0.89	0.88	0.85	0.87	0.80	**0.79**
	F1-score	0.59	0.61	0.63	0.63	0.63	**0.64**

word embeddings that are used represent all meanings of a word by the same vector, no matter how it is used. On the other hand, in the case of transformers, the word embedding represents the same words with different vectors depending on the context words. The transformers avoid recursion by processing the sentences as a whole rather than word by word and by learning relationships between words because of the multi-head attention mechanism and positional embeddings. The LSTM models are recurrent in nature and process each word of the sentence sequentially. With the increase in length of sentence, the runtime increases. But the transformer allows parallelization and reduces the training time.

It would also be relevant here to discuss the time complexity of the proposed model as compared to other models. For this purpose, the mean epoch times for different implementations are computed. Table 7 shows the mean epoch time of all implemented models. The reported mean epoch time of the proposed ensemble model is 391.3 s which is not much greater than other implemented models, particularly the transformer-based models. The slightly higher time requirement for the proposed model should be seen in respect to the performance evaluation metrics (refer to Tables 5 and 6). In case of the ensemble, all the models used in the ensemble are operating independently, indicating there is no dependency between them. All models are executed parallelly and are combined in the end using a simple voting scheme.

Finally, it is observed that the ensemble approach outperforms all other models. It may be due to the fact that the ensemble approach reduces the spread in the predictions made by the model. The ensemble model reduced variance, bias, and improved accuracy. Moreover, the ensemble model reduced the generalization error of the prediction. As compared to the top-performing individual models, the ensemble model improved the robustness or reliability of the average

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

Table 7. Category-wise distribution of tweets

Model	Mean Epoch time (in seconds)
CNN	44.0
LSTM	90.3
mBERT	367.2
XLM-RoBERTa	387.0
BETO	302.4
Ensemble	391.3

Table 8. Category-wise distribution of tweets

		Results					
		Subtask 1			Subtask 2		
S. No	Approach	P	R	F1-score	P	R	F1-score
1.	Transformer-based model (BETO,mBERT) (Turón *et al.* 2022)	–	–	77.32%	–	–	52.86%
2.	Transformer-based model (DistilBETO) (Tonja *et al.* 2022)	0.76	0.73	0.7455	0.49	0.49	0.4903
3.	Voting ensemble approach (Montañes-Salas *et al.* 2022)	–	–	0.76	–	–	0.50
4.	Ensemble model (Multi-tasking learning approach) (Vallejo-Aldana *et al.* 2022)	0.8032	0.7503	0.7759	0.47	0.47	0.4733
5.	Ensemble (CNN-Bi-LSTM, mBERT, BETO, XLM-RoBERTa (Proposed model)	**0.82**	**0.81**	**0.80**	**0.54**	**0.79**	**0.64**

performance of a model. This can be achieved by an ensemble model that minimizes noise, variance, and bias and enhances the stability and accuracy of stand-alone deep learning and transformer-based models.

5.2 Comparative analysis

The performance of the ensemble approach proposed is also compared with results obtained in some previous studies. Table 8 presents the performance measure values achieved by the previous studies as well as the ones obtained by the proposed ensemble model.

It may be noted that most of the previous studies experimented with transformer-based models, mainly BERT and BETO. BETO pretrained on the Spanish corpus was used to classify the tweets (Turón *et al.* 2022). The reported F1-score for both subtasks is 77.32 and 52.86, respectively. The other study used transfer learning and applied the transformer-based model (DistilBETO), which is the lighter version of the BETO (Tonja *et al.* 2022). The model achieved F1-score of 0.7455 for the first subtask and 0.4903 for the second subtask. The third and fourth studies used the ensemble approach to classify the tweets (Montañes-Salas *et al.* 2022; Vallejo-Aldana *et al.* 2022). While one used the voting ensemble approach, the other used an ensemble model involving a

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

multi-tasking learning approach. Using the voting ensemble approach, a F1-score of 0.76 was achieved for the binary classification subtask and a F1-score of 0.50 for multi-class classification. Subtask 1 achieved an F-1 score of 0.7759, while subtask 2 achieved an F-1 score of 0.4733 using the ensemble model. The proposed model is found to outperform all the previous implementations reported in various studies. It reports weighted average precision, recall, and F1-score of 0.82, 0.81, and 0.80, respectively, for the first subtask (binary classification). For the second subtask (multi-label classification), the proposed model obtains weighted average precision, recall, and F1-score of 0.54, 0.79, and 0.64, respectively. Thus, the obtained results of the proposed model are better for both subtasks – subtask 1 and subtask 2.

6. Conclusion

The paper reports application of some deep learning-based models (CNN and LSTM) and transformer-based models (mBERT, BETO, and XLM-RoBERTa) to classify the given set of tweets in Spanish as violent incidents or not. The second subtask of multi-label classification of tweets in different categories is also attempted. The paper proposes an ensemble of different models by combining them to achieve better performance. The proposed model for violent incident detection in the Spanish language performs better when compared to stand-alone or individual models as well as models proposed in the previous studies. Therefore, this research work adds to the knowledge pool of research in Spanish language in the form of a new approach which outperforms the existing methods. The model has been tested for both subtasks on violent incident detection, and better results are obtained in both the cases.

In this regard, it may be noted that mBERT and XLM-RoBERTa are multilingual models, whereas BETO is a monolingual model. The BETO model is pretrained exclusively on a large Spanish corpus. The multilingual models are pretrained on Wikipedia content and CommonCrawl data. The Spanish BERT (BETO) used a vocabulary of about 31K subwords using Byte Pair Encoding (BPE), while the multilingual BERT model has 119,547 tokens for all languages together (104 languages). Thus, the model with higher training vocabulary is likely to perform better. At the same time, diversity of training corpus is equally helpful. Thus, the better performance obtained on these models establish and strengthen the fact that language-specific model training is fundamental to achieving higher performance. The ensemble approach is found to be performing better than all the stand-alone models. This may be attributed to the fact that the constituent models are trained on different vocabulary, and therefore combining the models provides for a significantly large and heterogenous training. In this way, the ensemble model is able to combine the knowledge of individual models together for achieving better performance. Further, combining the models allows for fine-tuning and adjustment of various hyperparameters (learning rate, epsilon, and number of epochs), which also impact the model performance.

The applicability of the proposed model can also be explored in case of other languages if suitable datasets in those languages are available. As a future enhancement of this work, one may try to evaluate the model on other datasets and/or to further improve the performance by modifying the model. Further, the model can be provided with an additional module for fake incident reporting detection to make it more robust for processing different kinds of social media texts. In that way, the model would become more robust in the sense that it would be able to detect fake reported incidents and filter them out before labeling a reported incident as violent and classifying that into different categories provided.

Funding statement. This work is partly supported by HPE Aruba Centre for Research in Information Systems at BHU (No.: M-22-69 of BHU).

Competing interests. The authors declare that the manuscript complies with ethical standards of the journal, and there is no competing interests whatsoever.

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

References

Abdelfatah K. E., **Terejanu G.**, **Alhelbawy A. A.** (2017). Unsupervised detection of violent content in arabic social media. *Computer Science & Information Technology (CS & IT)*, 1–7. DOI: 10.5121/csit.2017.70401

ALSaif H. and Alotaibi T. (2019). Arabic text classification using feature-reduction techniques for detecting violence on social media. *International Journal of Advanced Computer Science and Applications* **10**(4), 77–87.

Annapragada A. V., **Donaruma-Kwoh M. M.**, **Annapragada A. V. and Starosolski Z. A.** (2021). A natural language processing and deep learning approach to identify child abuse from pediatric electronic medical records. *PLoS One* **16**(2), e0247404.

Arango A., **Pérez J.**, **Poblete B.**, **Proust V. and Saldaña M.** (2022). Multilingual resources for offensive language detection. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pp. 122–130.

Arellano L. J., **Escalante H. J.**, **Villaseñor Pineda L.**, **Montes y Gómez M. and Sanchez-Vega F.** (2022). Overview of da-vincis at iberlef 2022: Detection of aggressive and violent incidents from social media in spanish.

Basile V., **Bosco C.**, **Fersini E.**, **Nozza D.**, **Patti V.**, **Pardo F. M. R.**, **Rosso P. and Sanguinetti M.** (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 54–63.

Chen I. Y., **Alsentzer E.**, **Park H.**, **Thomas R.**, **Gosangi B.**, **Gujrathi R. and Khurana B.** (2020). Intimate partner violence and injury prediction from radiology reports. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*. World Scientific, pp. 55–66.

Conneau A., **Khandelwal K.**, **Goyal N.**, **Chaudhary V.**, **Wenzek G.**, **Guzmán F.**, **Grave E.**, **Ott M.**, **Zettlemoyer L. and Stoyanov V.** (2019). Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv: 1911.

Cumba-Armijos P., **Riofrío-Luzcando D.**, **Rodríguez-Arboleda V. and Carrión-Jumbo J.** (2022). Detecting cyberbullying in spanish texts through deep learning techniques. *International Journal of Data Mining, Modelling and Management* **14**(3), 234–247.

Devlin J., **Chang M.-W.**, **Lee K. and Toutanova K.** (2018). Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 1810.04805.

Díaz-Torres M. J., **Morán-Méndez P. A.**, **Villasenor-Pineda L.**, **Montes M.**, **Aguilera J. and Meneses-Lerín L.** (2020). Automatic detection of offensive language in social media: Defining linguistic criteria to build a mexican spanish dataset. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 132–136.

García-Díaz J. A., **Jiménez-Zafra S. M.**, **García-Cumbreras M. A. and Valencia-García R.** (2023). Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems* **9**(3), 2893–2914.

Hegarty K. and Tarzia L. (2019). Identification and management of domestic and sexual violence in primary care in the# metoo era: an update. *Current Psychiatry Reports* **21**(2), 1–8.

i Orts Ò.G. (2019). Multilingual detection of hate speech against immigrants and women in twitter at semeval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 460–463.

Khalafat M., **Ja'far S. A.**, **Al-Sayyed R.**, **Eshtay M. and Kobbaey T.** (2021). Violence detection over online social networks: an arabic sentiment analysis approach. *iJIM* **15**(14), 91.

Khatua A., **Cambria E. and Khatua A.** (2018). Sounds of silence breakers: Exploring sexual violence on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, pp. 397–400.

León-Paredes G. A., **Palomeque-León W. F.**, **Gallegos-Segovia P. L.**, **Vintimilla-Tapia P. E.**, **Bravo-Torres J. F.**, **Barbosa-Santillán L. I. and Paredes-Pinos M. M.** (2019). Presumptive detection of cyberbullying on twitter through natural language processing and machine learning in the spanish language. In *2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, IEEE, pp. 1–7.

Liu Y., **Ott M.**, **Goyal N.**, **Du J.**, **Joshi M.**, **Chen D.**, **Levy O.**, **Lewis M.**, **Zettlemoyer L. and Stoyanov V.** (2019). Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv: 1907.

Lopez I., **Quillivic R.**, **Evans H. and Arriaga R. I.** (2019). Denouncing sexual violence: A cross-language and cross-cultural analysis of# metoo and# balancetonporc. In *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2-6, 2019, Proceedings, Part II 17*, Springer, pp. 733–743

Mercado R. N. M., **Chuctaya H. F. C. and Gutierrez E. G. C.** (2018). Automatic cyberbullying detection in spanish-language social networks using sentiment analysis techniques. *International Journal of Advanced Computer Science and Applications* **9**(7), 228–235.

Montañes-Salas R. M., **del Hoyo-Alonso R. and Peña-Larena P.** (2022). Itainnova@ da-vincis: a tale of transformers and simple optimization techniques. In *IberLEF@ SEPLN*

Nobata C., **Tetreault J.**, **Thomas A.**, **Mehdad Y. and Chang Y.** (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 145–153.

Prasanth S., **Raj R. A.**, **Adhithan P.**, **Premjith B. and Kp S.** (2022). Cen-tamil@ dravidianlangtech-acl2022: Abusive comment detection in tamil using tf-idf and random kitchen sink algorithm. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pp. 70–74.

https://doi.org/10.1017/nlp.2024.25 Published online by Cambridge University Press

Qin G., **He J.**, **Bai Q.**, **Lin N.**, **Wang J.**, **Zhou K.**, **Zhou D. and Yang A.** (2022). Prompt based framework for violent event recognition in spanish. In *IberLEF@ SEPLN*

Ranasinghe T. and Zampieri M. (2021). Multilingual offensive language identification for low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing* **21**(1), 1–13.

Sharma D., **Gupta V. and Singh V. K.** (2024a). Abusive comment detection in tamil using deep learning. In *Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications*. Elsevier, pp. 207–226.

Sharma D., **Singh A. and Singh V. K.** (2024b). Thar-targeted hate speech against religion: A high-quality hindi-english code-mixed dataset with the application of deep learning models for automatic detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Subramani S., **Vu H. Q. and Wang H.** (2017). Intent classification using feature sets for domestic violence discourse on social media. In *2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*. IEEE, pp. 129–136.

Ta H. T., **Rahman A. B. S.**, **Najjar L. and Gelbukh A. F.** (2022a). Gan-bert: adversarial learning for detection of aggressive and violent incidents from social media. In *IberLEF@ SEPLN*

Ta H. T., **Rahman A. B. S.**, **Najjar L. and Gelbukh A. F.** (2022b). Multi-task learning for detection of aggressive and violent incidents from social media. In *IberLEF@ SEPLN*

Tonja A. L., **Arif M.**, **Kolesnikova O.**, **Gelbukh A. F. and Sidorov G.** (2022). Detection of aggressive and violent incidents from social media in spanish using pre-trained language model. In *IberLEF@ SEPLN*

Turón P., **Perez N.**, **Pablos A. G.**, **Zotova E. and Cuadros M.** (2022). Vicomtech at da-vincis: detection of aggressive and violent incidents from social media in spanish. In *IberLEF@ SEPLN*

Vallejo-Aldana D., **López-Monroy A. P. and Villatoro-Tello E.** (2022). Leveraging events sub-categories for violent-events detection in social media. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS. org*

Vaswani A., **Shazeer N.**, **Parmar N.**, **Uszkoreit J.**, **Jones L.**, **Gomez A. N.**, **Kaiser Ł. and Polosukhin I.** (2017). Attention is all you need. *Advances in Neural Information Processing Systems* **30**, 1–11.

Yu Y., **Si X.**, **Hu C. and Zhang J.** (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation* **31**(7), 1235–1270.

Cite this article: Sharma D, Gupta V, Singh VK and Pinto D. Should we stay silent on violence? An ensemble approach to detect violent incidents in Spanish social media texts. *Natural Language Processing* https://doi.org/10.1017/nlp.2024.25