DA-VINCIS at IberLEF 2023: Detecting Aggressive and Violent Incidents from Social Media in Spanish using Text Information

Ramón Zatarain Cabada[1], María Lucía Barrón Estrada[1], Víctor Manuel Bátiz Beltrán[1*], Ramón Alberto Camacho Sapien[1], Néstor Leyva López[1], Gerardo Ángel Beltrán Ruiz[1], Brandon Antonio Cárdenas Sainz[1] and Héctor Manuel Cárdenas López[1]

[1] *Tecnológico Nacional de México Campus Culiacán, Culiacán, Sinaloa, México*

Abstract

This article presents the work done in our participation in the task of detecting aggressive and violent incidents on Twitter by using images and text, in the DA-VINCIS competition as part of IberLEF 2023. Due to the high impact that a violent incident generates in society and the negative effect it has on the people involved in the event, finding a solution for the detection of such events is of vital importance for government institutions to ensure the safety of the population. Precisely, our participation was focused on making use of the corpus provided by the organizers to perform the task of detecting violent incidents using exclusively textual information. Different Natural Language Processing approaches were used to solve the competition task such as bag of words, textual representations, and transformers. Our proposal obtained, for subtask 1, an f1-score value of 0.9283 in the development phase and 0.8969 in the final phase, ranking second in the development phase and eighth in the final phase. For subtask 2, an f1-score value of 0.8380 was obtained in the development phase and 0.7647 in the final phase. The team's proposal ranked tenth in the development phase and thirteenth in the final phase.

Keywords

Aggressive Incidents Detection, Violent Incidents Detection, Text Classification, BERT

1. Introduction

Violence, defined as the action of causing physical or psychological harm or injury to others or to oneself, is a common problem that affects society around the world. On the other hand, violent incidents are a manifestation of violence and can be presented in a variety of ways, such as robbery, murder, harassment, terrorism, etc. For individuals and communities, experiencing an event of violence can imply the loss of a feeling of security [1] and that is why detecting these events is of major importance for government institutions to address and mitigate their effects by guaranteeing security to their population.

In this context, social networks as a popular medium of communication constitute an important part in the diffusion of violent incidents. More specifically, Twitter has served as a medium for sharing information about violent incidents that users of this social network witness or are aware of. This makes Twitter a useful tool for early detection of violent events due to its immediate and global nature. People can post real-time updates about what they are witnessing to inform authorities and other users about

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.
EMAIL: ramon.zc@culiacan.tecnm.mx (R. Zatarain); lucia.be@culiacan.tecnm.mx (M. L. Barrón); victor.bb@culiacan.tecnm.mx (V. M. Bátiz); ramon.cs@culiacan.tecnm.mx (R. A. Camacho); nestor.ll@culiacan.tecnm.mx (N. Leyva); 18170281@itculiacan.edu.mx (G. A. Beltrán); brandon.cs@culiacan.tecnm.mx (B. A. Cárdenas); hector.cl@culiacan.tecnm.mx (H. M. Cárdenas)
ORCID: 0000-0002-4524-3511 (R. Zatarain); 0000-0002-3856-9361 (M. L. Barrón); 0000-0003-4356-9793 (V. M. Bátiz); 0009-0003-9367-7730 (R. A. Camacho); 0000-0002-2767-5708 (N. Leyva); 0009-0000-1453-9064 (G. A. Beltrán); 0000-0001-9747-8534 (B. A. Cárdenas); 0000-0002-6823-4933 (H. M. Cárdenas)

violence in a specific location. This can be particularly beneficial in crisis or emergency situations, where decision making, and response coordination require accurate and up-to-date information.

Tweets containing information about violent events can be difficult to identify and classify, but there are some textual characteristics that can help differentiate these types of posts. The presence in the text of violent language, explicit descriptions of violence, mentions of victims or those involved, hashtags related to violence and references to specific places or events may suggest the presence of violent incidents.

In this paper, we present the work carried out in the participation in the DA-VINCIS competition as part of IberLEF 2023 [2] on the task of detecting aggressive and violent incidents on Twitter, particularly in text. For this, an analysis of the provided dataset was performed, identifying the distribution of the data, as well as the categorization of these. Also, a series of data cleaning and processing operations were applied to the corpus in order to eliminate non-relevant content in the text. Subsequently, several models for natural language processing were selected and evaluated, including models based on Transformers. As a product of all this work, the results and conclusions obtained are shown.

2. Related Work

Recognizing violent incidents on social media is strongly linked to understanding the context of the texts reporting these incidents, as reported in the work presented in Karystianis et. al. [3] where it is proposed to use police reports of domestic violence of various kinds (physical, psychological, emotional, and financial) as these are often recorded as long narratives. For this purpose, an approach based on syntactic text patterns was developed and evaluated on a set of police reports. Their findings suggest that the proposed methodology allows valuable information to be extracted to identify instances of violence (as previously mentioned) and the risk of these events escalating into more serious incidents. Hu et al. [4] proposed analyzing cases of conflict in the political domain using a pre-trained domain-specific language model using a variation of the BERT network, resulting in consistent data output within tests of 12 different datasets. The implementation of this type of algorithms in less controlled environments such as social networks was studied in Ha et al. [5] as the contents generated in these platforms are useful to define the prevalence, causes and consequences, by correlating them with the cases of unlabeled raw text violence, applying a machine learning model called DetectIPV, thus finding that this tool in the context of several applications achieves favorable outcomes in terms of detecting emotional and sexual abuse.

Prabhu et al. [6] use a BERT-based model tested in natural language processing (NLP) understanding. Such work explored the possibility of using BERT in combination with active learning strategies to label transaction descriptions, proving to be effective in the classification of multi-class text datasets. Another related work is reported in Piao [7] where text classification is done with a BERT transformer model called SciBERT in the school text classification task, where it was compared with other models and SciBERT was reported to have higher efficiency than other transformer networks. Finally, in the work by Liu et al. [8] a variation of BERT, called RoBERTa, was used to improve some previously ignored design decisions, resulting in better performance achieved by training the model for a much longer time.

In this work, a model based on BERT [9] was chosen to address the problem of violent event detection in social networks, since it is a model well suited to the problem of multi-class text classification, thus generating a model specifically trained for Subtask 1 of the competition and a model based on RoBERTa for Subtask 2, due to its strong performance in multi-class and multi-label problems.

3. Task Description

The DA-VINCIS competition [10] was divided into two subtasks: (1) identification of violent events and (2) recognition of categories of violent events.

Subtask 1: Identification of violent events. Determine whether a given tweet is associated with a violent incident from the considered categories or not (binary classification). In this two-class problem, the "positive" class consists of tweets reporting violent events from the categories of interest (accident, murder, and robbery). The negative samples are those tweets that are associated with reports of other violent events or no violent incidents.

Subtask 2: Violent event category recognition. Recognize the criminal category to which a given tweet belongs (multi-class multi-label classification). The categories considered are Accident, Murder, Robbery and Other. The category Other includes reports of violent incidents other than accident, murder, and robbery, as well as generic tweets not related to violent events.

For the evaluation of the subtask solution proposals, the competition organizers established that for subtask 1 the metrics of Accuracy, recall and f1-score would be used. Indicating that the primary metric would be f1-score. For subtask 2, the metrics used were Macro-average of Precision, recall and f1-score. Macro-average of f1-score was the primary evaluation metric for this subtask. The Codalab platform [11] was used for the submission of proposals and their evaluation.

4. Methodology

For the development of the competency, a methodology consisting of various stages was proposed: analysis of the data set, data preprocessing, model selection and training, model testing and proposal submission.

4.1. Dataset Description

The organizers of the competition provided a training dataset with 2996 records, each record containing two fields, one for the text of the tweet and another with the names of the images related to that text. A sample of the dataset is shown in Table 1. Additionally, two files containing the corresponding labels to each of the competition tasks were provided.

Table 1
Head of the dataset

Image Filenames	Text
['E_f1Fl-XMAMPFkT.jpg']	Morales: #EEUU “sufrió una derrota vergonzosa” ante el terrorismo y el narcotráfico #NarcoterrorismoDoméstico https://t.co/GCenoztH3H https://t.co/WMF36c68i3
['E8NYFP5WYAEpNRb.jpg']	Tus acciones te hacen ser una bella persona, no tu color de piel o tu dinero, esa joven con su hermoso gesto me robo el corazon. https://t.co/wXfw2Hrqv7
['E_BP33UXoAYQcTI.png']	#Seguridad 🚨 🔴 Detuvo policía municipal a dos hombres por robo a casa habitación, en #Puebla. En la acción se recuperaron computadoras portátiles, un equipo de audio, un dispositivo de comunicación móvil, entre otros artículos. Urbano| Noticias https://t.co/xXKSnVjz8k... https://t.co/HcoPbEpPLk
['E_XFlhCXIAAl-GR.jpg']	20 años del accidente que convirtió a Alex Zanardi en leyenda. https://t.co/TALvr20hB5
['E_Gc6blX0AEnhW5.jpg', 'E_Gc3IzWQA046Zf.jpg']	¡Llega el castigo para Max Verstappen! El neerlandés tendrá tres posiciones de sanción en la parrilla de salida del #RussianGP, tras el accidente que protagonizó con Lewis Hamilton en el #ItalianGP ¿Qué opinan de la decisión de la FIA? | #F1 https://t.co/yfUlLPzcGL

In subtask 1, we aimed to determine whether a given tweet is associated with a violent incident from the categories considered or not (binary classification). The label file contains the values 0 (negative class) or 1 (positive class) for each record in the training dataset. The 2996 records are divided into 1277 positive and 1719 negative (see left side of Figure 2).

In subtask 2, the aim is to recognize the incident categories to which a given tweet belongs (multi-class multi-label classification). The distribution of the tweets in the various categories can be seen on the right side of Figure 1.

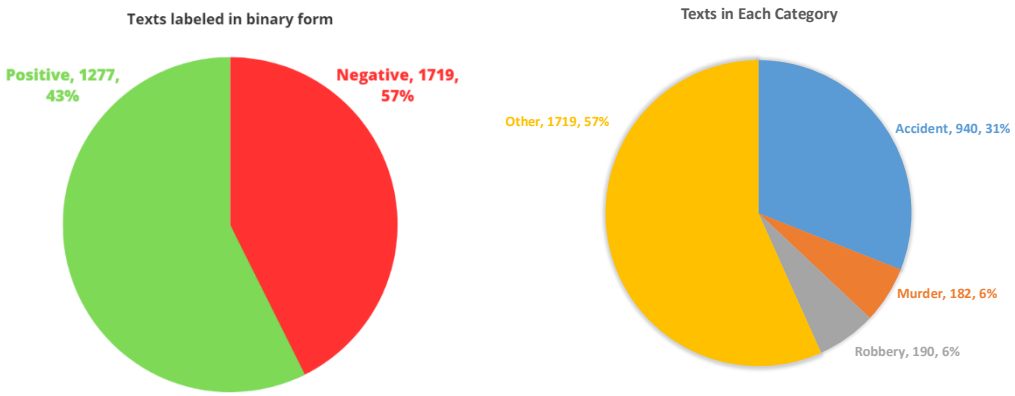

Figure 1: Dataset distribution whether the tweet is or not a Violent Incident (left) and by Categories (right).

For the evaluation of the proposed models, the organizers provided a dataset of 582 records in the development phase and for the final phase the test file had 1153 records.

4.2. Data Preprocessing

After the data analysis, the use of emojis, symbols, special characters, hashtags, and links was identified in the textual part of the dataset; therefore, the text was cleaned in order to perform a better training and detection. For the preprocessing of the data, the following activities were conducted:

- Remove symbology and special characters that do not provide context or information to the tweet.
- Convert the emojis present in the text to their textual representation in English using Python's "emoji" library. For example, for "👍" its textual representation would be: ":thumbs_up:".
- Normalize the different users included in the tweets. Users tagged in tweets do not directly contribute to the detection of violent incidents. For example, @News would be @User.
- Remove the links present in the tweet. Since in this case the detection of violent incidents is focused on text analysis, the links were removed to avoid noise in the data.

4.3. Model Selection and Training

At this stage, key decisions were made on the choice of classification models. It was decided to consider the following:

1. **Bag of Words (BoW):** This classical approach was used to represent the text of the tweets. This model consists of building a vocabulary from all the unique words present in the dataset, then representing each tweet as a feature vector indicating the frequency or presence of each word in the vocabulary. This is a simple but effective representation for training text classification models. The EvoMSA 2.0 library was used for this approach [12].
2. **Emoji Space from EvoMSA [12]:** The Emoji space textual model available in the EvoMSA 2.0 library was used to create a classification model, which was stacked to the bag-of-words-based model.
3. **Transformers based models:** For subtask 1, a model based on BERT (bert-base-multilingual-cased) was used. Such a model is pre-trained with information from the 104 languages with the highest content in Wikipedia using a masked language modeling (MLM) objective which was

introduced in [9]. The model is case sensitive. For subtask 2 we worked with a pre-trained model based on RoBERTa (roberta-base) [8]. This model is pretrained on the English language using a masked language modeling (MLM) objetive.

5. Results

For both phases of the competition (Development and Final), the data set provided was divided into 80% for training and 20% for validation. The results obtained in each phase and subtask are presented below.

5.1. Subtask 1

Initially, we used EvoMSA combined with bag-of-words. The proposed solution was built using the dataset given for validation and published on the Codalab platform for scoring. For the f1 metric, the proposal received a score of 0.9059. Later, it was determined to use the pre-trained model based on BERT (bert-base-multilingual-cased). This approach received a score of 0.9283 on the f1 measure, placing our proposal as second place in this phase of the competition, as shown in Table 2 (our submissions were made under the user VickBat but we will refer to the results under the team name ITC).

Table 2
Subtask 1 Development phase results

#	User	f1	Precision	Recall
1	Csuazob	0.9412 (1)	0.9372 (4)	0.9451 (1)
2	**ITC**	**0.9283 (2)**	**0.9283 (6)**	**0.9283 (4)**
3	Arnold	0.9283 (2)	0.9283 (6)	0.9283 (4)
4	EstebanPonce	0.9272 (3)	0.9139 (10)	0.9409 (2)
5	Mgraffg	0.9244 (4)	0.9469 (2)	0.9030 (6)
6	Agmegias	0.9231 (5)	0.9098 (11)	0.9367 (3)
7	alejandrov90	0.9224 (6)	0.9167 (9)	0.9283 (4)
8	Jorge	0.9191 (7)	0.9270 (8)	0.9114 (5)
9	HoracioJarquin	0.9142 (8)	0.9301 (5)	0.8987 (7)
10	danielvallejo237	0.9051 (9)	0.9491 (1)	0.8650 (10)
11	Rkcd	0.8989 (10)	0.9167 (9)	0.8819 (8)
12	bowlOfPetunias	0.8976 (11)	0.9279 (7)	0.8692 (9)
13	BrauuHdzm	0.8914 (12)	0.9393 (3)	0.8481 (11)
14	Darthremar	0.7905 (13)	0.8097 (12)	0.7722 (12)

For the final phase, work continued on the optimization of the pre-trained model based on BERT. In the final hyperparameters, we use a learning rate of 4e-5, AdamW as the optimizer, 128 as the maximum sequence value, and 8 as the training batch size. The solution proposal was generated with the data file provided to evaluate this phase. The submitted proposal obtained a value of 0.8969 in the f1-score metric and ranked eighth in the competition as can be seen in Table 3.

Table 3
Subtask 1 Final phase results

#	User	f1	Precision	Recall
1	danielvallejo237	0.9264 (1)	0.9302 (2)	0.9226 (5)
2	EstebanPonce	0.9203 (2)	0.9006 (8)	0.9409 (1)
3	Jorge	0.9186 (3)	0.9067 (5)	0.9308 (3)
4	agmegias	0.9165 (4)	0.8951 (11)	0.9389 (2)
5	csuazob	0.9100 (5)	0.8939 (12)	0.9267 (4)
6	Arnold	0.9069 (6)	0.9014 (7)	0.9124 (6)
7	rkcd	0.8991 (7)	0.9000 (9)	0.8982 (7)
8	**ITC**	**0.8969 (8)**	**0.9081 (4)**	**0.8859 (8)**
9	HoracioJarquin	0.8948 (9)	0.9456 (1)	0.8493 (12)
10	mgraffg	0.8903 (10)	0.9053 (6)	0.8758 (9)
11	escom	0.8822 (11)	0.8952 (10)	0.8697 (11)
12	BrauuHdzm	0.8822 (11)	0.8952 (10)	0.8697 (11)
13	Thisjesusalan	0.8693 (12)	0.8649 (14)	0.8737 (10)
14	d121201	0.8290 (13)	0.9183 (3)	0.7556 (14)
15	PabloGP	0.8251 (14)	0.8782 (13)	0.7780 (13)
16	pakapro	0.4639 (15)	0.4398 (15)	0.4908 (15)

5.2. Subtask 2

For this subtask, a model based on the pre-trained RoBERTa model was developed. In the development phase, the team's proposal obtained a result of 0.8375 in the Precision metric and 0.8380 in the f1 metric, ranking tenth in the stage. Table 4 shows the results of the stage.

Table 4
Subtask 2 Development phase results

#	User	f1	Precision	Recall
1	agmegias	0.8976 (1)	0.8796 (5)	0.9184 (1)
2	EstebanPonce	0.8968 (2)	0.8989 (3)	0.8948 (3)
3	alejandrov90	0.8960 (3)	0.9118 (1)	0.8817 (6)
4	Jorge	0.8960 (3)	0.9118 (1)	0.8817 (6)
5	Arnold	0.8859 (4)	0.8855 (4)	0.8862 (5)
6	csuazob	0.8847 (5)	0.8792 (6)	0.8907 (4)
7	HoracioJarquin	0.8706 (6)	0.8408 (8)	0.9040 (2)
8	rkcd	0.8445 (7)	0.8558 (7)	0.8344 (8)
9	danielvallejo237	0.8443 (8)	0.9038 (2)	0.8017 (9)
10	**ITC**	**0.8380 (9)**	**0.8375 (10)**	**0.8439 (7)**
11	BrauuHdzm	0.7890 (10)	0.8390 (9)	0.7663 (10)

For the final phase we worked on the optimization of the model by testing with various adjustments to the hyperparameters. In the final hyperparameters, we use a learning rate of 3e-5, AdamW as the optimizer, 512 as the maximum sequence value, and 8 as the training batch size. We sent the proposed solution to the test data set provided, and a result of 0.7760 was obtained in the Precision metric and

0.7647 under the f1 metric. The proposal was placed in the thirteenth position under the f1 metric. Table 5 shows the results of the subtask.

Table 5
Subtask 2 Final phase results

#	User	f1	Precision	Recall
1	EstebanPonce	0.8797 (1)	0.8737 (1)	0.8864 (3)
2	agmegias	0.8733 (2)	0.8523 (3)	0.8973 (2)
3	Jorge	0.8698 (3)	0.8622 (2)	0.8784 (4)
4	Arnold	0.8492 (4)	0.8305 (6)	0.8715 (5)
5	csuazob	0.8490 (5)	0.8441 (4)	0.8577 (6)
6	HoracioJarquin	0.8427 (6)	0.7663 (13)	0.9407 (1)
7	danielvallejo237	0.8421 (7)	0.8394 (5)	0.8449 (7)
8	BrauuHdzm	0.8048 (8)	0.8027 (9)	0.8091 (9)
9	escom	0.8036 (9)	0.8178 (8)	0.7974 (10)
10	Thisjesusalan	0.8030 (10)	0.7802 (11)	0.8294 (8)
11	devjesus	0.7789 (11)	0.8184 (7)	0.7517 (13)
12	rkcd	0.7773 (12)	0.7812 (10)	0.7781 (11)
13	**ITC**	**0.7647 (13)**	**0.7760 (12)**	**0.7571 (12)**
14	d121201	0.7116 (14)	0.7306 (15)	0.7210 (14)
15	PabloGP	0.6581 (15)	0.7338 (14)	0.6198 (15)
16	pakapro	0.2860 (16)	0.2531 (16)	0.4992 (16)

The results obtained show that both models based on traditional methods, such as the use of bag-of-words, and models based on newer techniques, such as the use of transformers (BERT and RoBERTa), were adequate for both tasks. The transformer-based models performed slightly better in our case, but we believe that overall the different models performed competitively.

6. Conclusions

In this paper, participation in the DA-VINCIS competition as part of IberLEF 2023 in the classification of violent incidents in tweets was presented. For subtask 1, the best result was obtained using a BERT-based pre-trained model with which the team's proposal placed second in the development stage and eighth in the final phase. For subtask 2, the best result was obtained using a pre-trained model based on RoBERTa with which the team's proposal ranked tenth in the development stage and thirteenth in the final phase. The results obtained were satisfactory, since a competitive level of performance was achieved in the competition tasks. As future work, it is proposed to continue with the optimization of the hyperparameters used in the classification models and to implement multimodal techniques to address the task.

7. Acknowledgements

We want to express our gratitude to CONAHCYT and the Tecnológico Nacional de México campus Culiacán for supporting our team to participate in the DAVINCIS@IberLEF 2023 challenge for detection of aggressive and violent incidents from social media in Spanish.

References

[1] Echeburúa, E., Corral, P. D., Amor, P. J. (2003). Evaluation of psychological harm in the victims of violent crime. Psychology in Spain, 7(1), 10-18.

[2] Jiménez-Zafra, S. M., Rangel, F., Montes-y-Gómez, M. (2023). Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, 2023.

[3] Karystianis, G., Adily, A., Schofield, P. W., Greenberg, D., Jorm, L., Nenadic, G., Butler, T. (2019). Automated analysis of domestic violence police reports to explore abuse types and victim injuries: Text mining study. Journal of Medical Internet Research, 21. doi:10.2196/13067

[4] Hu, Y., Hosseini, M., Parolin, E. S., Osorio, J., Khan, L., Brandt, P. T., D'orazio, V. J. (2022). ConfliBERT: A Pre-trained Language Model for Political Conflict and Violence (pp. 5469–5482). Retrieved from https://github.com/eventdata/

[5] Ha, P. T., D'silva, R., Chen, E., Koyutürk, M., Koyutürk, K., Unnur Karakurt, G. (2021). Identification of Intimate Partner Violence from Free Text Descriptions in Social Media. doi:10.1101/2021.12.15.21267694

[6] Prabhu, S., Mohamed, M., & Misra, H. (4 2021). Multi-class Text Classification using BERT-based Active Learning. Retrieved from http://arxiv.org/abs/2104.14289

[7] Piao, G. (2021). Scholarly Text Classification with Sentence BERT and Entity Embeddings. Retrieved from https://wikipedia2vec.github.io/wikipedia2vec/pretrained/

[8] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (7 2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. Retrieved from http://arxiv.org/abs/1907.11692

[9] Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Retrieved from https://github.com/tensorflow/tensor2tensor.

[10] Horacio Jarquín-Vásquez , Delia Irazú Hernández Farías, Joaquín Arellano, Hugo Jair Escalante, Luis Villaseñor-Pineda, Manuel Montes y Gómez, Fernando Sanchez-Vega. Overview of DA-VINCIS at IberLEF 2023: Detection of Aggressive and Violent Incidents from Social Media in Spanish. Procesamiento del Lenguaje Natural, Vol. XX, 2023.

[11] Pavao, A., Guyon, I., Letournel, A.-C., Baró, X., Escalante, H., Escalera, S., Thomas, T., Xu, Z. (2022). CodaLab Competitions: An open source platform to organize scientific challenges. Technical Report. Retrieved from https://hal.inria.fr/hal-03629462v1.

[12] Graff, M., Miranda-Jimenez, S., Tellez, E. S., Moctezuma, D. (2020). "EvoMSA: A Multilingual Evolutionary Approach for Sentiment Analysis [Application Notes]," in IEEE Computational Intelligence Magazine, vol. 15, no. 1, pp. 76-88, Feb. 2020, doi: 10.1109/MCI.2019.2954668.