

# Detection of Aggressive and Violent Incidents from Social Media in Spanish using Pre-trained Language Model

Atnafu Lambebo Tonja<sup>1</sup>, Muhammad Arif<sup>1</sup>, Olga Kolesnikova<sup>1</sup>, Alexander Gelbukh<sup>1</sup> and Grigori Sidorov<sup>1</sup>

<sup>1</sup>Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico

## Abstract

Violent and several other related problems, such as aggressive speech, offensive language, or bullying, are experiencing a growing online presence in the context of contemporary social media platforms. The research efforts towards detecting, isolating, and stopping these disturbing behaviors have intensified, in tight relation to the increasing performance of deep learning techniques applied in various Natural Language Processing (NLP) tasks. This paper presents the Instituto Politécnico Nacional, Centro de Investigación en Computación (CIC) team's system description paper for shared task @IberLEF2022. This study explores the applicability of language-specific pre-trained language model for tackling the problem of detection of aggressive and violent incidents from social media in Spanish for DA-VINCIS:@IberLEF2022 shared task. The proposed model on the DA-VINCIS dataset achieves F1 score of 0.7455 for violent event identification task (Task 1) and F1-score 0.4903 for violent event category recognition (Task 2).

## Keywords

Aggressive incidents, DistilBETO, Violent incidents, Spanish aggressive incident, Spanish violent incident, Social media,

## 1. Introduction

Online aggression is defined as any act of aggression, or a behavior intended to harm another person who does not wish to be harmed, that takes place using electronic media [1]. Nowadays, communication through social networks plays a significant role in social life. Social Networking Services (SSN) opens an entire universe of conceivable outcomes, yet they likewise address a significant threat, since users are exposed to many threats and attacks; among them violent comments, which can cause short term and long-term damage victims [2]. Social media companies made various attempts toward detecting, removing, and stopping these behaviors, with both Twitter and Facebook rolling out several tools for flagging and reporting unwanted pieces of content [3] however, these efforts encountered several problems. First, a very small percentage

---

IberLEF 2022, September 2022, A Coruña, Spain.


✉ atnafu.lambebo@wsu.edu.et (A. L. Tonja); mariff2021@cic.ipn.mx (M. Arif); kolesolga@gmail.com (O. Kolesnikova); gelbukh@cic.ipn.mx (A. Gelbukh); sidorov@cic.ipn.mx (G. Sidorov)

🌐 <https://atnafuatx.github.io/> (A. L. Tonja)

🆔 0000-0002-3501-5136 (A. L. Tonja); 0000-0001-6141-0204 (M. Arif)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

of the victims even consider using these tools [4, 5], thus, they are unaware of the offences towards them. Second, the amount of data that needs to be flagged and analyzed by human moderators is enormous. For example, [6] estimates that 6,000 tweets are posted online every second. These reasons powered the increasing research efforts toward automated processes, grounded in natural language processing (NLP) techniques, for the identification and removal of aggressive, offensive, or hateful online social media content.

Nevertheless, creating an explained corpus of social media content reasonable for this work ended up being an exceptionally challenging errand because of the emotional and fluctuating definitions of the labels [7]. To promote research in identifying aggressive and violent incidents in Spanish, DA-VINCIS at IberLEF2022 arranged a task to track such social media incidents in Spanish tweets [8]. In this paper, we explore the feasibility of automatically determining if a text obtained from Twitter describes a violent incident or not by using a language-specific pre-trained language model for the shared task at DA-VINCIS@IberLEF2022. The paper is organized as follows: section 2 describes related work section 3 gives an overview of the dataset statistics and the description of the task, section 4 explains the methodology adopted in this work, section 5 emphasizes the experimental results and analyzes them. Finally, section 6 concludes the paper and sheds some light on possible future work.

## 2. Literature Review

In recent years, the automatic detection of aggressive behavior in social media is gaining a lot of attention. Early approaches relied heavily on feature engineering combined with traditional machine learning classifiers such as naive bayes and support vector machines [9]. However, machine learning algorithms evolved somewhat recently, with numerous NLP systems being developed and employed for such problems [10]. Their approach aims to improve the performance for detection and interpretation of sentiments and opinions in texts, by employing semantic web and artificial intelligence techniques [5]. For the first time, a violent speech detection corpus together applying with several machines learning algorithms (logistic regression, random forest, among others) and pre-processing techniques (TF-IDF scores, stemming) was described in [5].

The authors [6] also analyzed the corpus from [6] and tried to improve the results using skip-gram features. Studies which improved the results on the same corpus by employing a convolutional neural network (CNN) as classifier were [11, 12]. Also, [13] analyzed both CNNs and gated recurrent units (GRU) [14] achieving better results. Concerning aggressiveness detection related work, [15] classifies Facebook comments using three deep learning architectures namely, CNN, Long Short Term Memory (LSTM) networks, and Bi-directional Long Short Term Memory (Bi-LSTM) networks. The authors also combine these architectures with a majority voting-based ensemble method. Another study[16] briefly explained the properties of social media bullies and violent aggressors and found that stalkers post with less frequency, participate in fewer online communities, and are less popular than users with standard models of behavior. Their research shows that machine learning algorithms can accurately detect users who exhibit bullying and aggressive behavior, with more than 90% of accuracy.

Pre-trained Language Models (PLMs) are large neural networks that are used in a wide variety of NLP tasks [17]. Models are first pre-trained over a large text corpus and then fine tuned on a

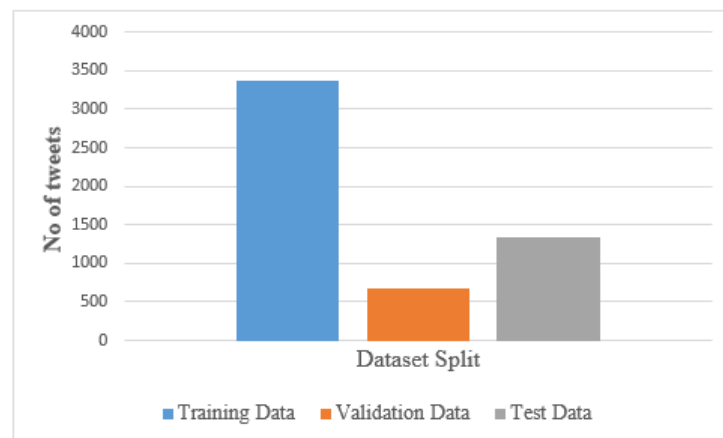
downstream task. PLMs are thought of as good language encoders, supplying basic language understanding capabilities that can be used with ease for many downstream tasks [17, 18]. There are a lot of PLMs trained using monolingual corpus of certain languages. Some of these are :-

- Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for NLP pre-training developed by Google [19]. Trained by English monolingual corpus.
- Spanish BERTa (BETO)-BERT-based language model pre-trained exclusively on big Spanish corpus [20].
- ALBETO and DistilBETO - which are versions of ALBERT and DistilBERT pre-trained exclusively on Spanish corpora, ALBETO ranging from 5M to 223M parameters and one of DistilBETO with 67M parameters [21]

In this paper we explore the effect of using language-specific PLMs for detecting aggressive and violent incidents. We used DistilBETO to test the applicability and performance on detecting aggressive and violent incidents in Spanish text. We selected these PLMs because (1) It is pre-trained exclusively on Spanish corpora, (2) It is a lightweight version of BETO that can be easily fine-tuned with a small amount of compute resource.

### 3. Data Description

In the experimentation phase, we used DA-VINCIS corpus compiled from Twitter by retrieving tweets associated with violent incidents. The corpus contains train, validation, and test sets of Spanish Tweeter data. Figure 1 shows the dataset distribution used in this task. The training

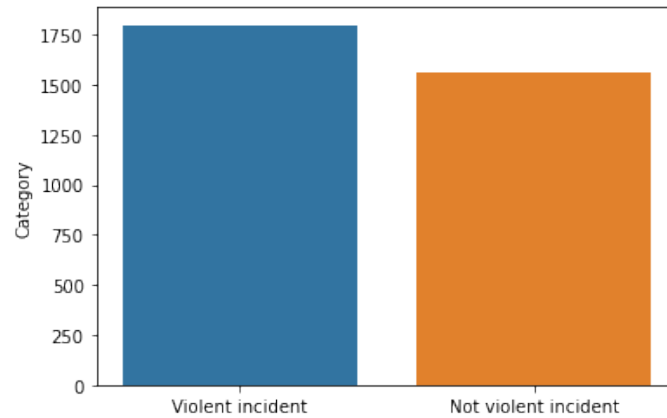


**Figure 1:** Dataset distribution of training, validation and test data

data includes 3,362 tweets with different labels for two sub tasks.

The aim of this task is to build models that are able to determine if a news item describes a violent incident or not by analyzing textual information. The shared task has two sub-tasks. The first sub-task (Task 3a) is violent event identification :-

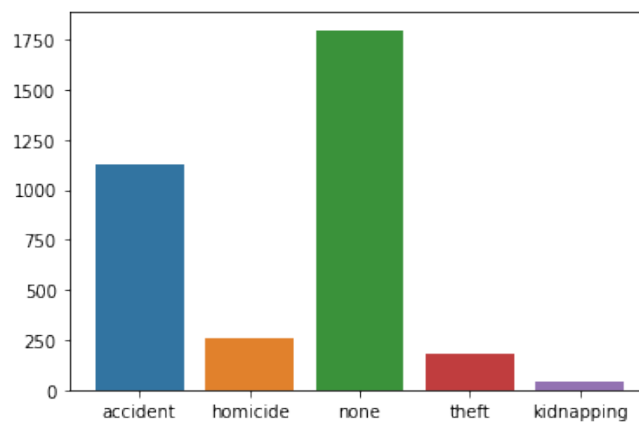
- Determining whether a given tweet is associated with a violent incident or not, this is a binary classification task. As shown in Figure 2, this task has training data with two labels violent incident and not violent incident.



**Figure 2:** Sub Task 1 data distribution with labels

The second sub-task is violent event category recognition:-

- Recognizing the crime category to which a given tweet belongs, this is a multi-class classification task. As shown in Figure 3 this task has training data with five labels:- accident (accidente), homicide (homicidio), theft (robo), kidnapping (secuestro), and none (ninguno).



**Figure 3:** Sub Task 2 data distribution with labels

### 3.1. Data Pre-processing

The corpus collected from Twitter include unwanted data and needs further pre-processing before the experiment step. In order to clean the datasets we removed URLs, emojis and hashtags. We also removed unwanted characters using the python String library that includes punctuation.

## 4. Proposed Approach

To explore the applicability and effect of using language-specific PLMs for detection of aggressive and violent incidents in social media for Spanish language we used distilled version of BETO (DistilBETO). DistilBETO - is one of is a small, fast, cheap and light Transformer model based on the BETO architecture trained with 67M parameters [21]. After pre-processing the textual data as described in section 3 then, the text sequence is tokenized using the subword tokenizer included with the distill-bert-base-spanish model with maximum text length of 128. For fine-tuning we used transformers library from huggingface (<https://huggingface.co/>). To optimize the model we used Adam optimizer with a batch size of 64 and learning rate of 0.0001. We used the maximum number of epochs of 10 with earlystopping based on the performance of the validation set. We also used dropout of 0.2 to regularize the model. For run our experiment we used Google colab pro + with python programming language. Figure 1 summarizes the parameters used to fine-tune distilled version of BETO pre-trained model.

Parameters	Values
Dropout	0.2
learning_rate	0.0001
epochs	10
optimizer	adam
max_length	128
batch_size	64

**Table 1**  
Parameters used in the proposed approach

## 5. Experiments and Result

After using the DistilBETO pre-trained model described in section 4, we obtained the results shown in Table 2 and 3 for violent event identification task (Task 1) and for violent event category recognition (Task 2) respectively. In both tables we presented our results with the other teams result that have better performance according to the result published by the organizers [8], the the results are ordered according to the values of F1-score in the descending order.

As shown in Table 2 our model performed with 76, 73, and 74.55 scores for precision, recall, and F1-score respectively in the violent event identification task (Task 1). Table 2 using a language-specific pre-trained model shows the competitive results compared to the result of other teams for violent event identification tasks.

Task	Team	Precision	Recall	F1-score
Task 1	LuisArellano	0.82	0.82	0.7817
	danielvallejo237	0.80	0.75	0.7759
	sdamian	0.76	0.75	0.7562
	pturon	0.82	0.70	0.7555
	Bernardo	0.78	0.73	0.7548
	ITAINNOVA	0.79	0.72	0.7529
	Abu	0.76	0.74	0.748
	<b>CIC-atnafu</b>	<b>0.76</b>	<b>0.73</b>	<b>0.7455</b>

**Table 2**  
Results of Task 1 and comparison with other teams

For violent event category recognition (Task 2), as shown in Table 3, our pre-trained model performed, with 49, 49, and 49.03 scores for precision, recall, and F1-score respectively. As compared to the result of other teams using language-specific models like DistilBETO which is trained using Spanish corpus can give a competitive result in violent event category recognition tasks.

Task	Team	Precision	Recall	F1-score
Task 2	kelven	0.55	0.55	0.5543
	pturon	0.53	0.53	0.5286
	ITAINNOVA	0.5	0.5	0.5046
	LuisArellano	0.46	0.57	0.4981
	<b>CIC-atnafu</b>	<b>0.49</b>	<b>0.49</b>	<b>0.4903</b>

**Table 3**  
Results of Task 2 and comparison with other teams

The result demonstrates that language-specific pre-trained language models can give promising results for violent event identification and categorization task in Spanish.

## 6. Conclusion

In this paper, we explored the application of language-specific pre-trained language models to detect incidents from social media for Spanish. A distilled version of BETO pre-trained model has shown a promising result, and our team is achieved 8th place for Task 1 and 5th for Task 2. In the future, we will explore the performance of other language-specific pre-trained models on the same task and how a bigger amount of data can influence the performance of pre-trained models.

## Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1S-47854 of CONACYT, Mexico, grants 20220852, 20220859, and 20221627 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the

CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

## References

- [1] C. N. DeWall, C. A. Anderson, B. J. Bushman, The general aggression model: Theoretical extensions to violence., *Psychology of Violence* 1 (2011) 245.
- [2] G. Ortiz, H. Gómez-Adorno, J. Reyes-Magaña, G. Bel-Enguix, G. Sierra, Detection of aggressive tweets in mexican spanish using multiple features with parameter optimization., in: *IberLEF@ SEPLN*, 2019, pp. 520–525.
- [3] M.-A. Tanase, D.-C. Cercel, C.-G. Chiru, Upb at semeval-2020 task 12: Multilingual offensive language detection on social media by fine-tuning a variety of bert-based models, *arXiv preprint arXiv:2010.13609* (2020).
- [4] E. Greevy, A. F. Smeaton, Classifying racist texts using a support vector machine, in: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, pp. 468–469.
- [5] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017, pp. 512–515.
- [6] S. Malmasi, M. Zampieri, Challenges in discriminating profanity from hate speech, *Journal of Experimental & Theoretical Artificial Intelligence* 30 (2018) 187–202.
- [7] Z. Waseem, T. Davidson, D. Warmsley, I. Weber, Understanding abuse: A typology of abusive language detection subtasks, *arXiv preprint arXiv:1705.09899* (2017).
- [8] L. V.-P. M. M. y. G. F. S.-V. Luis Joaquín Arellano, Hugo Jair Escalante, Overview of da-vincis at iberlef 2022: Detection of aggressive and violent incidents from social media in spanish., in: *IberLEF@ SEPLN*, volume 69, September 2022.
- [9] U. Arbieu, K. Helsper, M. Dadvar, T. Mueller, A. Niamir, Natural language processing as a tool to evaluate emotions in conservation conflicts, *Biological Conservation* 256 (2021) 109030.
- [10] E. Cambria, P. Chandra, A. Sharma, A. Hussain, Do not feel the trolls, *ISWC*, Shanghai (2010).
- [11] B. Gambäck, U. K. Sikdar, Using convolutional neural networks to classify hate-speech, in: *Proceedings of the first workshop on abusive language online*, 2017, pp. 85–90.
- [12] V. Peñaloza, Detecting aggressiveness in mexican spanish tweets with lstm+ gru and lstm+ cnn architectures., in: *IberLEF@ SEPLN*, 2020, pp. 280–286.
- [13] Z. Zhang, D. Robinson, J. Tepper, Detecting hate speech on twitter using a convolution-gru based deep neural network, in: *European semantic web conference*, Springer, 2018, pp. 745–760.
- [14] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, *arXiv preprint arXiv:1412.3555* (2014).
- [15] S. Madisetty, M. S. Desarkar, Aggression detection in social media using deep neural

- networks, in: Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), 2018, pp. 120–127.
- [16] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, A. Vakali, Mean birds: Detecting aggression and bullying on twitter, in: Proceedings of the 2017 ACM on web science conference, 2017, pp. 13–22.
  - [17] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, Y. Goldberg, Measuring and improving consistency in pretrained language models, Transactions of the Association for Computational Linguistics 9 (2021) 1012–1031.
  - [18] R. Weng, H. Yu, S. Huang, S. Cheng, W. Luo, Acquiring knowledge from pre-trained model to neural machine translation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 9266–9273. doi:10.1609/aaai.v34i05.6465.
  - [19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
  - [20] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.
  - [21] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, Albeto and distilbeto: Lightweight spanish language models, arXiv preprint arXiv:2204.09145 (2022).