# GAN-BERT: Adversarial Learning for Detection of Aggressive and Violent Incidents from Social Media

Hoang Thang Ta[1,2], Abu Bakar Siddiqur Rahman[1,3], Lotfollah Najjar[3,*] and Alexander Gelbukh[1]

[1]*Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico*
[2]*Dalat University, Lam Dong, Vietnam*
[3]*College of Information Science and Technology, University of Nebraska Omaha, Omaha, Nebraska, USA*

**Abstract**

In this paper, we address Subtask 1 of Detection of Aggressive and Violent INCIdents from Social Media in Spanish (DA-VINCIS). We introduced our method, using text embeddings from pre-trained transformer models for the training process by GAN-BERT, an adversarial learning architecture. Finally, we obtained F1 of 74.43%, Precision of 74.08%, and Recall of 74.79% on Subtask 1.

**Keywords**

Offensive Language, Violence Detection, GAN-BERT, Text Classification, NLP, DA-VINCIS, IberLEF

## 1. Introduction

Violence can be defined as an aggressive behavior that is used intentionally by showing power, physical force, threatening attitude of a group of people or individuals to the oppressed group of people or individuals. Society is adversely affected by this notorious activity that creates psychological disorder, depression, anxiety, post-traumatic stress disorder both for the people who experienced and witnessed it [1]. Violence is a component of terrorism that coherent with the official United States (U.S.) definition of domestic terrorism [2]. Millions of United States residents are affected by interpersonal violence, intimate partner violence, sexual violence that impede economic development [3]. Violence can be done by homicide, robbery, or kidnapping. As a solution, a government must know the violence activities in real time to provide the population's security. Recently, social media such as Twitter and Facebook are platforms that can be used to detect the violence or monitor violence activities from users who post suspicious related texts during a certain period. To detect the violence from social media, Natural Language Processing (NLP) researchers used machine learning and deep learning techniques to search for effective solutions to these issues.

*IberLEF 2022, September 2022, A Coruña, Spain.*

*Corresponding author.
tahoangthang@gmail.com (H. T. Ta); abubakarsiddiqurra@unomaha.edu (A. B. S. Rahman); lnajjar@unomaha.edu (L. Najjar); gelbukh@cic.ipn.mx (A. Gelbukh)
0000-0003-0321-5106 (H. T. Ta); 0000-0002-8581-0891 (A. B. S. Rahman); 0000-0003-3960-4189 (L. Najjar); 0000-0001-7845-9039 (A. Gelbukh)

NLP was used to identify violent presence in armed actors in Colombia by parsing the text based on an identification from a Spanish corpus that was collected human rights violation for three decades consisting of the date, description, location of the violation and the identifier of each narrative who described about the violation [4]. Social media consists of different violent content as text or topics. Violent detection models can able to identify the texts that are related to violent content and extract the violent topics from social media [5]. By nature, humans can get mental illness with physical violence. There needs to be identification from clinical texts to know the statistics about physical victimisation. A machine learning based free text based NLP algorithm was proposed in [6] by concerning mental disorders because of physical violence. Due to the diversity of different languages, it is hard to identify violent content from text. In Arabic sentences, it is possible to use violent words in a non violent Arabic sentence. K-means can not separate both violent and non violent due to high dimensional space. To cluster in a low dimensional space, a sparse Gaussian process latent variable model can be used in Arabic social media to separate violent content from non violent content [7].

In this paper, we participate in Subtask 1 of the DA-VINCIS workshop on IberLEF 2022 [1] to detect aggressive and violent incidents from social media. There are 2 subtasks in total, in the form of binary classification and multi-label classification. The first subtask (Subtask 1) is to detect whether a given tweet has violent incidents or not. The second task (Subtask 2) will clarify which one of 5 categories of violent events that a given tweet belongs to. We apply some preprocessing steps on tweets before putting them to the training process by GAN-BERT [8]. Different from the original architecture, noises were modified before feeding to the generator network. They have a random rate and the same size of the hidden layer of transformers. We also set up different runs, from the data addition in the same domain to using back translation to increase the training data.

## 2. Related Work

Violent behavior critically imperil the security of individuals, society and for a country. This motivates to detect violent events to mitigate the violent activities. Social media is the first step where people post about unexpected concurrent incidents of the environment that can unlock an option to identify the violent activities by automatically monitoring the high amount of violent related keywords on the posts of a location for a certain period of time.

NLP researchers can also contribute to national discussion by creating a gun violence database (GVDB) [9] to improve the social science research by using information retrieval, coreference resolution, and event detection. The database contained 7366 articles with annotation where the articles were collected from 1512 cities of the United States (U.S.). The articles contained information about the incidents of gun violence, some articles also provided information about shooter or victim identity and weapon on the incidents. The database helps to identify the victims and the people who created violence. Arellano et. al. collected a Spanish corpus from twitter that contained 3412 tweets related to violent topics [10]. Osorio et. al. generated a geo-referenced database to track the violence presence in Mexican criminal Organization (MCO) between 2000 to 2008 [11]. The tracking of violence in the database was implemented in four

[1] https://codalab.lisn.upsaclay.fr/competitions/2638

steps by scraping the web, the classification of news stories based on machine learning classifiers, event coding and data visualization by time series and geographic information systems. Event coding used a supervised NLP application for extracting events known as Eventus ID [12]. In the database, Eventus ID comprises the code of the actor who is conducting an action, the location and date.

Violence can occur in different aspects of society such as school violence, youth violence as gang, inter partner violence, domestic violence that causes a serious deleterious effects on society by heath damage of the people, loosing their money for the victim of robbery, injured or lost their loved ones to oppressed by kidnapping or homicides. School violence has impacted the entire school environment from students, professors to stuff. The violence causes dropout rates in the school, creates non productivity of the students, less attendance in the school. An interview was taken for 131 students from 39 schools in the United States (U.S.) to predict the risk of school violence [13]. Key Linguistic features were extracted from the interview based on NLP technologies and support vector machine, logistic regression, artificial neural network classifiers were used to predict a binary class classification based on risk prediction from the school violence. Gang violence is another threat to society. Firearm violence increased by 40% from 2014 to 2015 in Chicago, a state of the United States (U.S.) by more than 3000 victims of shootings. Patton et. al. used a digital urban violence analysis approach and used NLP methods to build an automatic classifier to classify tweets as aggression, grief, others. 800 tweets were collected that were posted by Chicago gang members and some young participants from Chicago. The goal of the task was to build a NLP tool to predict the clusters of aggression and loss for the youth involved in gangs in Chicago [14]. Inter partner violence (IPV) creates instability between the members of a family with strongly negative effects on the children. IPV increased more during the COVID-19 pandemic. 6348 annotated tweets were collected by Al-Garadi et. al. where BERT and ROBERTa methods were used to identify IPV and non IPV tweets [15]. Some other works have done on IPV where contexual word embedding, Tf-IDF, Word2Vec, machine learning algorithms such as SVM, MLP, Random Forest, Logistic Regression, and Naive Bayes were used [16, 17]. Lin et. al. used a new method BERT as sequence tagging task to identify risks on gun violence from social media data, GVDB where BERT used as an embedding layer and BIO sequence tagging method used for word decoding similar to token portions of BERT. Then LSTM, BiLSTM and CRF methods are methods to classify each token to identify the risks of gun violence [18].

Domestic and sexual violence (DSV) effects more with severe mental illness (SMI) patients compared to general sound health people. According to statistics 27% women and 13% men with SMI experienced DSV, the amount is 9% and 5% for women and men respectively in the case of the general population [19]. A NLP model was used to extract interpersonal violence from clinical texts of electronic health records. BioBERT is a model for fine-tuning the annotated dataset and the model was evaluated by 10 fold cross validation [20]. Mensa et. al. proposed a violence episodes detection systems based on semantic information of texts to detect violence episodes in emergency room reports that contains 150k annotated dataset from an European Union injury database projects with the two annotations of violence and non-violence related injuries [21].

Based on the literature reviews on violence detection from social media, most of the work has used BERT, ROBERTa, BioBERT, machine learning classifiers. There is no work related to

**Table 1**
Some examples in the new training set. The tweets were passed by some pre-processing steps.

| tweet | vio. | acci. | homi. | non-vio. | rob. | kidnap. |
|---|---|---|---|---|---|---|
| Patrullero muere en accidente en la circunvalar vía @usuario | 1 | 1 | 0 | 0 | 0 | 0 |
| Breaking , primera imagen de Hamilton luego de su accidente . @usuario | 0 | 0 | 0 | 1 | 0 | 0 |
| Seis policías serán sancionados por su conducta durante asalto al Capitolio | 0 | 0 | 0 | 1 | 0 | 0 |
| Siguen investigaciones por el asesinato de un patrullero en Malambo | 1 | 1 | 0 | 0 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... |

GAN-BERT to detect violent events and categorizing the topics of violent events. This motivates us to experience GAN BERT to detect and classify the violence events from Twitter data.

## 3. Task Description

As a task in IberLEF 2022, DA-VINCIS (Detection of Aggressive and Violent INCIdents from Social Media in Spanish) aims to identify violent events from users' tweets, which were collected from Twitter [10]. Participants are required to classify these tweets into correct categories in 2 types, binary classification and multi-label classification. This is correspondingly to 2 subtasks, violent event identification and violent event category recognition. According to organizers, this is the first edition of the task, so it is only in Spanish. The task data was packed in DA-VINCIS corpus for both tasks, and participants can join on one of the subtasks or both.

In Subtask 1, from a given tweet, a classifier is needed to define whether this tweet has a violent incident or not. This subtask is considered as a problem of binary classification, and contains only 1 category (`Violent`). Subtask 2 involves identifying a violent event category of a given tweet, considered as a problem of multi-label classification. A tweet can belong to many categories depending on the content it conveys. There are 5 event categories: `Accident`, `Homicide`, `Non-Violent-incident`, `Robbery`, and `Kidnapping`.

## 4. Dataset Analysis

From the challenge website, we download 3 groups of datasets: training set, trial set, and validation set [10]. The original training and trial sets contain 3 files for each: data file, label file for Subtask 1, and label file for Subtask 2. We merge original training and trial sets, including their files into a new training set, shown in Table 1 by some examples. For the remaining content of this paper, we mention this new training set as the training set.

The training set contains 3412 tweets, classified in 6 categories: `Violent`, `Accident`, `Homicide`, `Non-violent-incident`, `Robbery`, and `Kidnapping`. The category `Violent` is an inverse version of the category `Non-violent-incident`.

Since we participate in Subtask 1, we only care about the `Violent` category. This category has 1587 tweets containing violent incidents (label = 1) and 1825 for non-violent ones (label = 0), accounting for 46.51% and 53.48% correspondingly.

The organizers also provided a test set with 1344 tweets and the validation set with 673 tweets, and they both have no labels. We came to the challenge late so we could not be able to test our method on the validation set. However, we can add them as unlabeled data for the training by GAN-BERT.

## 5. Methodology

Firstly, we apply some preprocessing steps on tweets to remove redundant content, which are assumed not to contribute much for the model performance. We also normalize some content that helps the model learn better.

- Remove special characters, smileys, and symbols.
- Remove urls starting with http:// or https://, such as `https://t.co/BNkgYv7a1B`. We observe that all urls in the dataset are from the domain `t.co`.
- Normalize hashtags by removing `#`. For example, hashtags `#SOSUSA`, `#CerroAzul`, `#Violencia` will be normalized as `SOSUSA`, `CerroAzul`, and `Violencia`.
- Normalize `@[user]` to `@usuario`.
- Apply package `es_core_news_md` of spaCy v2.3.2[2] to split tweets into tokens and remove redundant spaces, then combine them back as texts.

To increase the data amount, we apply a method of data augmentation, back translation. We used pre-trained Marian models of Helsinki-NLP in Hugging Face[3] to translate original tweets in Spanish to English, French, German, and Italian. For a given tweet, 2 texts were collected in the back translation, its translated text and its back translation text. We did this only on the training set for all violent categories. Then, the `Violent` category has 13520 tweets with violent incidents (label = 1, distribution = 46.41%), and 15606 with non-violent ones (label = 0, distribution = 53.58%).

In the training process, we apply `GAN-BERT` [8], BERT-like architecture with unlabeled data in a generative adversarial setting. According to the authors, `GAN-BERT` is appropriate for small datasets when we can not collect more data because of expensive costs and time consuming. Figure 1 describes the architecture of `GAN-BERT`. Traditionally, SS-GAN (Semi-Supervised Generative Adversarial Networks) include 2 networks: (1) a discriminator $D$ for classifying input dataset, (2) a generator $G$ to differentiate fake data in an adversarial manner. In GAN-BERT, these networks are both multilayer perceptron networks. Furthermore, BERT is added on the top of SS-GAN to provide sentence embeddings and contextualized embeddings of the words in a sentence. In GAN-BERT, both generator and discriminator are Multi Layer Perceptron networks (MLP).

The real data including 2 categories (labeled and unlabeled) will be passed to BERT to receive text embeddings. For example, the training set is labeled data because its tweets were already

[2]https://spacy.io/

[3]https://huggingface.co/Helsinki-NLP

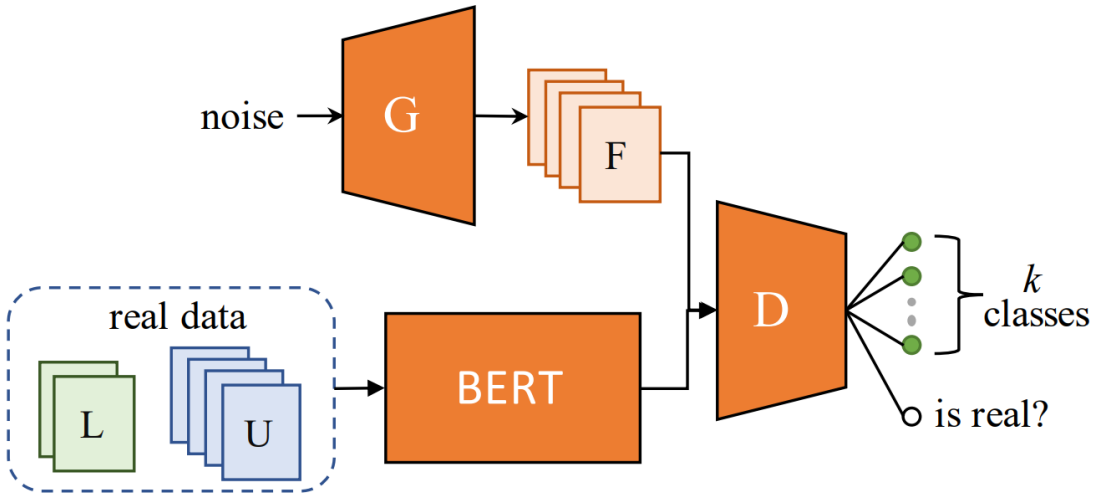


**Figure 1:** GAN-BERT architecture: A set of fake sentence pairs F, given a random distribution, will be generated by G. The discriminator D will take these fake pairs, unlabeled U and labeled L vector representations computed by BERT as input for the training process [8].

labeled (0 for no violence, and 1 for violence). The test set and the validation are unlabeled data because their data is unknown. In this paper, because there have only 2 categories (0 and 1), we set violent tweets as labeled data, otherwise data (including non-violent tweets) is unlabeled. Meanwhile, noises or fake tweets will be generated based on real text embeddings (vectors). The original paper created noises as 100-dimensional vectors drawn from a normal distribution *N(0, 1)*. However, we modified this noise method. When real examples were generated from BERT, we took their hidden layer embeddings (768-dimensional), generated randomly a noise rate (from 0 to 1), then distorted the real data by this rate with normal distribution *N(-1, 1)*. We believe that this way makes noises closer to the real data. After that, the discriminator D will learn and differentiate between real and fake tweets. If a tweet is real, it will be classified to one of $k$ classes (in this case only 1 for violence). In contrast, this tweet will be classified to class $k + 1$ if it is fake. To be clear, the output of discriminator is a multiclass classifier, in which class $k + 1$ is fake and classes from $1$ to $k$ are real classes. In our paper, there is only a real class, which contains violent tweets.

## 6. Experiment

We apply several ways to add extra data in the same domain, including test set, validation set, and back translation set, as shown in Table 2. This practice is expected to improve the inner representation of GAN-BERT as mentioned in Section 5. After the data addition, we filter out the repetitive tweets if they occur in the training set.

Table 2 shows our 4 runs with different configurations and Precision results. Except for the third run with training on full data, we split the training set into new training and validation sets with the ratio 8:2. The first run used `learning_rate=2e-7` and `epoch=20` while the others

**Table 2**
Our results in Subtask 1 from organizers' website.

| # | Additional data | Data splitting | Learning rate | Epoch | Precision |
|---|---|---|---|---|---|
| 1 | **test_set** | **8:2** | **2e-7** | **20** | **0.7408** |
| 2 | test_set + val_set | 8:2 | 2e-5 | 40 | 0.6944 |
| 3 | test_set + val_set | full | 2e-5 | 40 | 0.6592 |
| 4 | back_translation + test_set + val_set | 8:2 | 2e-5 | 40 | 0.7216 |

go with `learning_rate=2e-5` and `epoch=40`. In the third run, the model could be overfitting when training on the full data. In this case, there is no doubt that the model performance is the worst with Precision of 65.92%. The second run and the fourth run had better results, with Precision values of 69.44% and 72.16% respectively. Meanwhile, we obtained the best results in the first run, with Precision of 74.08%. According to the final results of organizers, this run also had F1 of 74.43%, and Recall of 74.79%.

It is still not clear about the correlations of additional data, data splitting, learning rate and number of epochs used in the experiment. In theory, the fourth run should have the best precision due to the largest amount of additional data it has in the training process. Compared to the second run on the same configuration, we can prove this theory is correct. However, since the first and third runs apply different configurations, we are unable to compare them with the second and the fourth runs. Due to the submission limit, we could not be able to clarify this by producing more test sets.

## 7. Conclusion

In this paper, we engaged Subtask 1 of the DA-VINCIS challenge (detect aggressive and violent incidents in tweets) and applied GAN-BERT to solve the problem. Before the training process, we used some preprocessing steps to handle tweets, and back translation to improve the number of tweets in the `Violent` category. In the experiment, we tried on 4 runs with different configurations, and obtained the best values: F1 of 74.43%, Precision of 74.08%, and Recall of 74.79%.

In the experiment, we are unable to clarify the role of back translation and other set ups such as data splitting, learning rate, and the number of training epochs in GAN-BERT. Therefore, we will continue with extra work on these to see the correlations between them in the future. Besides, we also prefer to apply other methods based on adversarial networks for a better performance for the task and relevant ones.

## Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20220852 and 20220859 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 also show high gratitude to Holland computing center, University of Nebraska to provide their high computing GPU resources. The authors acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

## References

[1] E. G. Krug, J. A. Mercy, L. L. Dahlberg, A. B. Zwi, The world report on violence and health., in: The lancet 360, 9339, 2002, pp. 1083–1088.

[2] S. Jones, The evolution of domestic terrorism., in: tatement before the House Judiciary Subcommittee on Crime, Terrorism, and Homeland Security CSIS-Center for Strategic and International Studies Washington, DC, 2022.

[3] S. A. Sumner, J. A. Mercy, L. L. Dahlberg, S. D. Hillis, J. Klevens, D. Houry, Violence in the united states: status, challenges, and opportunities., in: Jama 314, 5, 2015, pp. 478–488.

[4] J. Osorio, M. Mohamed, V. Pavon, B.-O. Susan, Mapping violent presence of armed actors in colombia., in: Advances of Cartography and GIScience of the International Cartographic Association, volume 16, 2019, pp. 1–9.

[5] C. Basave, A. Elizabeth, Y. He, K. Liu, J. Zhao, A weakly supervised bayesian model for violence detection in social media., in: Sixth International Joint Conference on Natural Language Processing: Proceedings of the Main Conference, Asian Federation of Natural Language Processing, 2013, pp. 109–117.

[6] V. Bhavsar, J. Sanyal, R. Patel, H. Shetty, S. Velupillai, R. Stewart, M. Broadbent, J. H. MacCabe, J. Das-Munshi, L. M. Howard, The association between neighbourhood characteristics and physical victimisation in men and women with mental disorders., in: BJPsych open 6, 4, 2020.

[7] K. E. Abdelfatah, G. Terejanu, A. A. Alhelbawy, Unsupervised detection of violent content in arabic social media., in: Computer Science Information Technology (CS IT) 9, 2017.

[8] D. Croce, G. Castellucci, R. Basili, Gan-bert: Generative adversarial learning for robust text classification with a bunch of labeled examples, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 2114–2119.

[9] E. Pavlick, H. Ji, X. Pan, C. Callison-Burch, The gun violence database: A new task and data set for nlp., in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, p. 1018–1024.

[10] L. J. Arellano, H. J. Escalante, L. Villaseñor-Pineda, M. Montes-y Gómez, F. Sanchez-Vega, Overview of DA-VINCIS at IberLEF 2022: Detection of Aggressive and Violent Incidents from Social Media in Spanish., in: SEPLN journal, volume 69, Septiembre 2022.

[11] J. Osorio, A. Beltran, Enhancing the detection of criminal organizations in mexico using ml and nlp., in: International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–7.

[12] J. Osorio, A. Reyes, Supervised event coding from text written in spanish: Introducing eventus id., in: Social Science Computer Review 35, 3, 2017, pp. 406–416.

[13] Y. Ni, D. Barzman, A. Bachtel, M. Griffey, A. Osborn, M. Sorter, Finding warning markers:

leveraging natural language processing and machine learning technologies to detect risk of school violence., in: International journal of medical informatics, 139, 2020.

[14] D. U. Patton, K. McKeown, O. Rambow, J. Macbeth, Using natural language processing and qualitative analysis to intervene in gang violence: A collaboration between social work researchers and data scientists., in: arXiv preprint arXiv:1609.08779, 2016.

[15] M. Al-Garadi, A. Sarker, Y. Guo, E. Warren, Y.-C. Yang, S. kim, 134 automatic identification of intimate partner violence victims from social media., in: BMJ journals, 2022.

[16] T. Yallico Arias, J. Fabian, Automatic detection of levels of intimate partner violence against women with natural language processing using machine learning and deep learning techniques., in: Annual International Conference on Information Management and Big Data, Springer, Cham, 2022, pp. 189–205.

[17] I. Y. Chen, E. Alsentzer, H. Park, R. Thomas, B. Gosangi, R. Gujrathi, B. Khurana, Intimate partner violence and injury prediction from radiology reports., in: BIOCOMPUTING 2021: Proceedings of the Pacific Symposium, 2020, pp. 55–66.

[18] H.-Y. Lin, T.-S. Moh, B. Westlake, Gun violence news information retrieval using bert as sequence tagging task., in: 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 2525–2531.

[19] H. Khalifeh, P. Moran, R. Borschmann, K. Dean, C. Hart, J. Hogg, D. Osborn, S. Johnson, L. M. Howard, Domestic and sexual violence against patients with severe mental illness., in: Psychological medicine 45, 4, 2015, pp. 875–886.

[20] R. Botelle, V. Bhavsar, G. Kadra-Scalzo, A. Mascio, M. V. Williams, A. Roberts, S. Velupillai, R. Stewart, Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: an applied evaluative study., in: MJ open 12, 2, 2022.

[21] E. Mensa, D. Colla, M. Dalmasso, M. Giustini, C. Mamo, A. Pitidis, D. P. Radicioni, Violence detection explanation via semantic roles embeddings., in: BMC medical informatics and decision making 20, 1, 2020, pp. 1–13.