

# Multi-Task Learning for Detection of Aggressive and Violent Incidents from Social Media

Hoang Thang Ta<sup>1,2</sup>, Abu Bakar Siddiquir Rahman<sup>1,3</sup>, Lotfollah Najjar<sup>3,\*</sup> and Alexander Gelbukh<sup>1</sup>

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

<sup>2</sup>*Dalat University, Lam Dong, Vietnam*

<sup>3</sup>*College of Information Science and Technology, University of Nebraska Omaha, Omaha, Nebraska, USA*

## Abstract

In this paper, we participate in the task of Detection of Aggressive and Violent INCIDENTS from Social Media in Spanish (DA-VINCIS). We apply a multi-task learning network, MT-DNN to train users' tweets on their text embeddings from pre-trained transformer models. In the first subtask, we obtained the best F1 of 74.80%, Precision of 75.52%, and Recall of 74.09%. Meanwhile, F1 of 39.20%, Precision of 37.79%, and Recall of 43.88% are results in the second subtask.

## Keywords

Violence Detection, Multi-task learning, MT-DNN, Text Classification, DA-VINCIS, IberLEF

## 1. Introduction

Violence activities intend to use power, physical torture or threatening behavior by an individual or a community against a powerless person [1]. This act is not only inimical for an individual's life temporarily but also the victims experienced mentally injurious for a long period of time. Post violence effects are responsible for psychological disorders, depression, anxiety, fear of normal life, creating instability, and removing trust from society. These results have a detrimental impact both for the witnesses and victims of the violence. Homicide, robbery, kidnapping are some categories of violence. Violence can be classified in different segments of human relationship and society such as interpersonal violence, intimate partner violence, sexual violence, youth violence. Millions of United states residents are adversely affected by these violations each year [2]. The government has an important role to prevent the violations and provide safety for the people of the country. In this modern era of online networking services through social media, it is easy to pass violence activities by texting each other. Moreover, social media contributes to spreading any news about crisis or violence rapidly compared to

---

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

\*Corresponding author.

✉ tahoangthang@gmail.com (H. T. Ta); abubakarsiddiquirra@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)



© 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)

news media. The 2011 Japan earthquake are the examples for faster communication by social media [3]. Therefore, to prevent the violations, it would be helpful to identify the violence activities from social media. Machine learning and deep learning techniques are used by Natural Language Processing (NLP) researchers to analyze the texts from user posts in Facebook, Twitter or other social media to identify the violence.

Mental illnesses are related to physical violence. By analyzing the clinical texts, it would be easier to identify physical victimisation by developing and evaluating a machine learning based NLP algorithm [4]. The annotation of the data is one of the challenging tasks after collecting a dataset. Violence detection model (VDM) can identify violent related words and violent related topics from social media data without having the annotation of data by only knowing whether a particular word in the text indicates violence or not with the knowledge of the VDM model [5]. It is quite hard to find an available violence related dataset from social media and an appropriate text based method to identify violence in text due to lack of proper resources and distinct structure of grammar in different languages. Arellano et. al. collected a Spanish dataset from Twitter about violence events and described an overview of shared tasks of aggressive and violent events detection (DA-VINCIS [6]) on IberLEF 2022. Abdelfatah et. al. found that the same Arabic words exist in both violent and non violent texts. The better results were achieved to separate the violent and non violent tweets with the clustering of low dimensional latent space of sparse Gaussian process latent variable model (SGPLVM) whereas the identification of both violent and non violent tweets were not possible by using k-means in high dimensional space [7]. SGPLVM can work without the annotation of data. Therefore, methods chosen are an important task about the identification of violent events. Kotze et. al. collected the messages from 26 WhatsApp groups that share the texts related to protest and violence in South Africa. Logistic regression classifier with unigram and Word2Vec feature models performance better to detect the violence incidents from the texts in the WhatsApp messages [8].

In this paper, we used MT-DNN (Multi-Task Deep Neural Networks) to identify violent incidents from tweets by using text embeddings from pre-trained transformers [9]. We attended both subtasks of the DA-VINCIS challenge 2022 (<https://codalab.lisn.upsaclay.fr/competitions/2638>). Except this section, Section 2 introduces related works and popular methods to detect aggressive and violence events in texts. Section 3 and Section 4 show the tasks in detail and analyze the datasets provided by organizers. Our methodology used to train the model is presented in Section 5. Lastly, we report our results by some experiments, as well as withdraw conclusions and declare our future works, in Sections 6 and Section 7 correspondingly.

## 2. Related Works

Violent behavior lessens people's security in the society. This necessitates identifying violent behavior from the social media data. These unexpected activities can be noticed in a country by different types of violence from individuals to every aspect of society. Gun violence, inter partner violence, domestic violence, school violence, youth violence are some types of violence that causes instability.

Pavlick et. al. proposed a gun violence database that can assist social science researchers and the government to identify the victims and who afflicted the violence. 7366 articles were collected

from 1512 cities of the United States (U.S.) to create the database that provides information about the incidents of gun violence, the shooter or the victim demographic and weapons that were used in the violence [10]. The risk of gun violence impacts the victims. Lin et. al proposed a new method to use BERT as a sequence tagging task based on Information retrieval. The method identifies the risks of gun violence with the gun violence database [11]. Intimate partner violence often creates a situation where women feel more insecure and do not speak up. Arias et. al. detected psychological violence against women by using 5 annotations of psychological levels in 5250 Spanish records that were collected from 6 different sources. Psychological experts performed the 5 data labelling based on the text of the reports. The labelling indicates about the steps of victims level by Low Risk, Emotional Blackmail, Jealousy/Justification, Insults/Humiliations, and Threats/Possessiveness. TF-IDF and Word2Vec were used for the word vectors and the detection of the risks experimented by five machine learning classifier SVM, MLP, Random Forest, Logistic Regression, and Naive Bayes and two deep learning methods LSTM and BiLSTM [12]. Transformer based NLP methods BioBERT is able to identify the reference of interpersonal violence from clinical data in electronic health records. However, the methods are not able to identify the reason for preventing violence. This research is useful because of knowing the most dangerous sources of violence that affects the health damage [13]. School violence is one of the alarming situations for the nation that results in increasing dropout rates, insecurity for the students to attend in the class, psychological illness that diminish the scholar mindset among the students. Ni et. al. found a dataset based on the interview of 131 students from 39 schools to predict the risk of school violence based on support vector machine, logistic regression, artificial neural network classifiers [14].

Violence-related datasets are usually collected from past news articles or Twitter posts from specific locations with known violent activities. Osorio et. al. collected a Spanish corpus based on human right violation for three decades from 1988 to 2017 for mapping the violence presence of armed actors in the Colombian Civil war. The corpus provides the information of the date, location, description and witness of the violation. NLP techniques and geographic information systems were applied for early warning detection systems to detect emerging, intense and critical situations [15]. Gang violence is the most threatening for the people of a country. Patton et. al. collected 800 tweets from twitter that were posted by Chicago gang members. NLP methods were used to build an automatic classifier to classify tweets based on aggression, grief, others and for predicting the clusters of aggression and loss for the youth who were involved in gangs. 3000 victims were shot by firearm violence that increased about 40% in Chicago during a year from 2014 to 2015 [16]. A Geo referenced database was generated for tracking the violence of the Mexican Criminal Organization from 2000 to 2008 [17]. Al-Garadi et. al. collected 6348 annotated tweets to identify inter partner violence by using BERT and RoBERTa [18].

Based on all the methods in literature review and according to our best knowledge, no work has been done by using multi task deep neural networks (MT-DNN) for identifying violent incidents. MT-DNN consists of two layers where the lower layers share all tasks and the top layer used for task specific purposes. In the lower layer section, each word is represented as an embedding vector and then the embedding vector fed to a transformer encoder section to generate the contextual embedding vectors. The top layer is used for task specific text classification purposes.

### 3. Task Description

The DA-VINCIS (Detection of Aggressive and Violent INCIDENTs from Social Media in Spanish) challenge aims to detect aggressive and violent events from users' tweets, which were gathered from Twitter. There are 2 subtasks (Subtask 1 and Subtask 2), violent event identification and violent event category recognition, correspondingly in the form of binary classification and multi-label classification. From a given tweet, a classifier is required to detect the correct categories that this tweet belongs to. This is the first edition of DA-VINCIS so organizers only offer tweets in Spanish. Hopefully, the challenge will be upgraded with more languages in the future. DA-VINCIS corpus contains all data for both subtasks, and participants can join in one of them or both.

In Subtask 1, from a given tweet, a classifier is needed to identify whether this tweet has a violent incident or not. This subtask is viewed as a problem of binary classification, and has only 1 category (Violent). For example, "La cárcel de madre e hijo por presunto secuestro de un joven discapacitado en Meta." is a violent tweet, while "Alcalde de Guaimaca se salva de morir en accidente" is not.

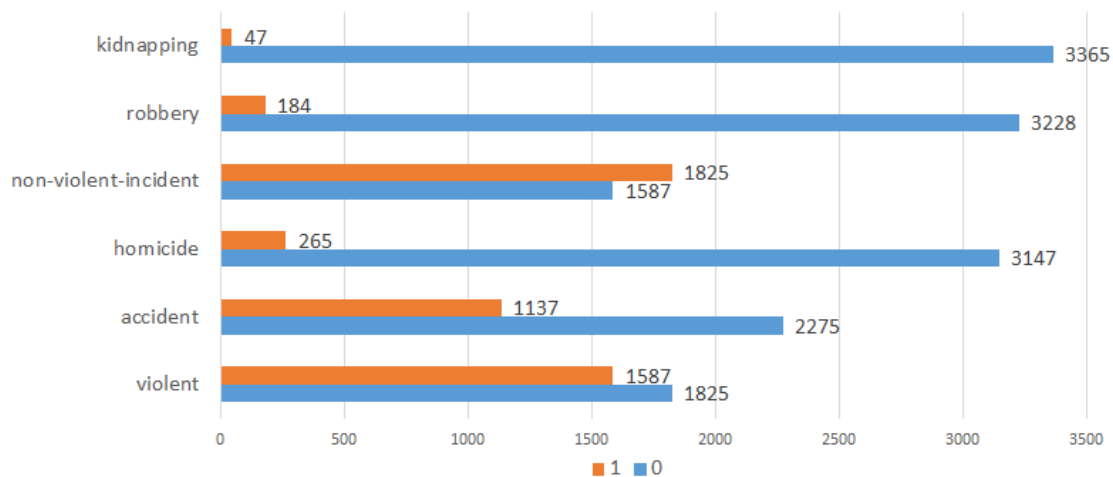
Subtask 2 involves detecting a violent event category of a given tweet, considered as a problem of multi-label classification. A tweet can belong to many categories depending on the content it conveys. There are 5 event categories: Accident, Homicide, Non-Violent-incident, Robbery, and Kidnapping. In the previous example, "La cárcel de madre e hijo por presunto secuestro de un joven discapacitado en Meta." also contains a kidnapping event.

### 4. Dataset Analysis

There are 3 groups of datasets: training set, trial set, and validation set. The original training and trial sets have 3 separate files: data file, label file for Subtask 1, and label file for Subtask 2. We mix these sets (original training and trial sets, including their files) into a new training set. For the remaining content of this paper, we mention this new training set as the training data. Later on, the organizers also published a test set with 1344 tweets and the validation set with 673 tweets, and they both have no labels, only raw tweets.

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. We use both categories to avoid some mistakes happening when getting data for the training process in Subtask 1 and Subtask 2.

Figure 1 shows the distribution of tweets by violent categories in our new training set. Here, we combine Subtask 1 and Subtask 2 together, so in total the data has 6 categories. We have 2 labels (0 and 1) for a tweet that belongs to a certain category or not. We found that the distribution of Kidnapping, Robbery, and Homicide categories is highly imbalanced. Therefore, we decided to apply back translation to increase the number of tweets in these categories. The more detail will be presented in Section 5.



**Figure 1:** The distribution of tweets by violent categories in the training set. Label 1 for a tweet belongs to a category and 0 if it does not belong to that category.

## 5. Methodology

To handle the training data, we use some preprocessing steps, which remove or normalize unwanted content from tweets and may not contribute much for the model performance.

- 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`.
- Convert `@[user]` to `@usuario`.
- Use package `es_core_news_md` of `spaCy v2.3.2` (<https://spacy.io/>) to parse tweets into tokens and discard redundant spaces, then combine tokens 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 (<https://huggingface.co/Helsinki-NLP>) 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 apply 2 methods:

- Apply back translation for imbalanced categories (Kidnapping, Robbery, and Homicide) as mentioned in Section 4 and Figure 1. Then, the new results and the original data was combined in a file named `data_augmented1.csv`.
- Apply back translation for categories. We did the same thing as the previous method and the new file is `data_augmented2.csv`.

Table 1 shows the distribution of violent categories by datasets. The distribution of `data_augmented1` is more balanced than the other two. However, we are not sure that the more

**Table 1**

The distribution of datasets after using back translation.

	original_data		data_augmented1		data_augmented2	
	0	1	0	1	0	1
accident	2275	1137	5739	1196	19556	9570
homicide	3147	265	4604	2331	26821	2305
non-violent-incident	1587	1825	5188	1747	13520	15606
robbery	3228	184	5285	1650	27496	1630
kidnapping	3365	47	6514	421	28708	418

balanced data can guarantee a better model performance. Because we have a multilingual dataset when using back translation, we decided to use a pre-trained model, bert-base-multilingual-cased for getting text embeddings in the training.

Finally, we use MT-DNN, a Multi-Task Deep Neural Network to learn representations among multiple natural language understanding (NLU) tasks. This network can be applied to solve some problems, such as using-sentence classification (COLA, SST-2), pairwise text similarity (STS-B), pairwise text classification (MPRC, QQP), and pairwise ranking (QLNI) by tasks together. The top layers generate more special representations to produce different outputs by tasks, while the lower layers are shared embeddings in the training process [9]. Instead of using 2 classifiers for 2 subtasks, we work only on Subtask 2, and infer the Violent category as the inverse results of the Non-violent-incident category. For MT-DNN, we create 5 tasks based on 5 violent categories in the form of binary classification, and create their \*.json data for the training process.

## 6. Experiment

Every team has 5 submissions for each subtask, and our team only used 4 with various configurations, shown in Table 2. We set the same learning rate  $1r=2e-5$  for all runs. In the first run, our model was trained on data\_augmented1 with full data and in 10 epochs. The fourth run is the same as the first one, but it was trained in 20 epochs. The second run worked on data\_augmented2 with data splitting into training and validation sets with the ratio 8:2 and 20 epochs. The third run was also trained on data\_augmented2, but without data splitting. It is the best run, with Precision values of 75.52% and 39.20% corresponding to Subtask 1 and Subtask 2. In this third run, we also obtained the best F1 of 74.80% and Recall of 74.09% for Subtask 1, and F1 of 39.20% and Recall of 43.88% for Subtask 2, according to organizers.

In the experiment, we still are not clear about the correlations of model configurations, especially the role of data splitting. Since we apply MT-DNN only on the available data offered by organizers instead of combining other tasks with their own data, our performance is not the best one compared to other teams. This network allows us to train models very fast with less epochs compared to traditional transformer models such as BERT.

**Table 2**

Our precision results in Subtask 1 and Subtask 2 from organizers' website.

#	Training data	Data splitting	Learning rate	Epoch	Subtask 1	Subtask 2
1	data_augmented1	full	2e-5	10	0.7296	0.3750
2	data_augmented2	8:2	2e-5	20	0.7264	0.3864
3	<b>data_augmented2</b>	<b>full</b>	<b>2e-5</b>	<b>20</b>	<b>0.7552</b>	<b>0.3920</b>
4	data_augmented1	full	2e-5	20	0.7312	0.3789

## 7. Conclusion

In this paper, we engaged tasks of the DA-VINCIS challenge and applied MT-DNN, a multitask learning network, to identify aggressive and violent events in tweets. Before the training process, we used some preprocessing steps to remove or normalize unwanted components in tweets and back translation in 4 languages to make the distribution of categories more balanced. We designed our training with 5 subtasks in the form of binary classification, corresponding to 5 violent categories. Then, there were 4 runs with different configurations for the training. We obtained the best F1 of 74.80%, Precision of 75.52%, and Recall of 74.09% in Subtask 1, while F1 of 39.20%, Precision of 37.79%, and Recall of 43.88% are results in Subtask 2.

In the future, we will continue to research about the architecture of MT-DNN and multi-task learning to improve the results. We also want to know more about the correlations of model configurations, especially data splitting and learning rate to see how it affects the model performance when using MT-DNN.

## 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] The world report on violence and health., author=Krug, Etienne G. and Mercy, James A. and Dahlberg, Linda L. and Zwi, Anthony B., in: The lancet 360, 9339, 2002, pp. 1083–1088.
- [2] S. A. Sumner, J. A. Mercy, L. L. Dahlberg, S. D. Hillis, J. Kleven, D. Houry, Violence in the United States: status, challenges, and opportunities., in: Jama 314, 5, 2015, pp. 478–488.
- [3] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event

- detection by social sensors., in: Proceedings of the 19th international conference on World wide web, 2010, pp. 851–860.
- [4] 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.
  - [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] 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.
  - [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] E. Kotzé, B. A. Senekal, W. Daelemans, Automatic classification of social media reports on violent incidents in South Africa using machine learning., in: *South African Journal of Science* 116, 3-4, 2020, pp. 1–8.
  - [9] X. Liu, P. He, W. Chen, J. Gao, Multi-task deep neural networks for natural language understanding, *arXiv preprint arXiv:1901.11504* (2019).
  - [10] 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.
  - [11] 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.
  - [12] 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.
  - [13] 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.
  - [14] 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.
  - [15] 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.
  - [16] 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.
  - [17] 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.

- [18] 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.