CICESE at DA-VINCIS 2023: Violent Events Detection in Twitter using Data Augmentation Techniques

Esteban Ponce-León[1,*], Irvin Hussein López-Nava[1]

[1]*Department of Computer Science, Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California (CICESE), México.*

Abstract

This paper describes our participation in the shared evaluation campaign of DA-VINCIS at IberLEF 2023. In this work, we address the subtasks proposed, Violent Event Identification (subtask 1) and Violent Event Category Recognition (subtask 2) using multimodal information from tweets (text and images), by using a Bidirectional Encoder Representations from Transformers (BERT) with and without data augmentation techniques. For text augmentation, the GPT-3 model and prompt engineering were used meanwhile for image augmentation an image recovery approach from the web was used, and image captioning to handle the images from the visual information. Our approach obtained second place for subtask 1 (F1 = 0.9203) and first place for subtask 2 (F1 = 0.8797) among 16 different teams.

Keywords

Violence Detection, Social Media, Data Augmentation, Natural Language Processing, BERT, Image Captioning

1. Introduction

Currently, the use of social networks has changed the way in which information is shared and is now a relevant part of government communication agencies and companies [1]. The information collected from social networks is a valuable input to analyze the flow of information, opinions, and feelings [2]. For example, in recent years, there has been a growing interest in online social media monitoring and big data analytics. Social media platforms are used to collect information and, in some cases, to examine the prediction of crime [3, 4]. However, much of the research to date has only focused on US cities, and therefore, on publications written in English [5]. One of the efforts to promote research in Spanish language related to this topic is DA-VINCIS [6], a task from the shared evaluation campaign of Natural Language Processing systems in Spanish and other Iberian languages (IberLEF). This task is framed in the field of event classification using information extracted from Twitter's post. There are several application scenarios, for example, in some countries it is difficult to measure the crime incidence rate accurately due to the lack of trust that the population has towards the authorities, causing insufficient data and, in the worst case, leading to conclusions that are far from reality. For example, in Mexico City 9 out of 10 crimes are not reported, where the lack of trust towards the authorities is accentuated,

IberLEF 2023, September 2023, Jaén, Spain

[*] Corresponding author.

esteban@cicese.edu.mx (E. Ponce-León); hussein@cicese.mx (I. H. López-Nava)

https://sites.google.com/view/husseinlopeznava (I. H. López-Nava)

and only 1 out of 100 reported cases reach a sentence [5]. As a consequence, in recent years, local and national newscasts have been taking advantage of social media to promote citizen complaints and help authorities take actions in neglected areas or duties.

In this paper, we describe our methodology to approach the task presented on DA-VINCIS IberLEF 2023, which consists of the detection of violent incidents on Twitter using both images and text. Our main contribution consisted on the exploration of data augmentation with the help of existing Large Language Models (LLM), based on the current dataset provided, which has relatively small data for training and with a high imbalanced in the different categories, as can be seen in Section 2. At the same time, the use of LLM allowed us to have a new kind of data that differs from the original more than in word level. Therefore, we addressed the subtasks from the event by using a Bidirectional Encoder Representations from Transformers (BERT), data augmentation through Generative Pre-trained Transformer (GPT), and a Bootstrapping Language-Image Pre-training (BLIP) for image captioning to process the visual information in order to improve the performance of the model.

The rest of the paper is organized as follows: Section 2 describes the purpose of the shared tasks and provides some basic statistics from the dataset. Section 3 explains the methodology of our proposed solution. Section 4 details our experiment setup as well as the final results. Finally, in Section 5 our main conclusions are presented.

2. Data and Task Description

DA-VINCIS is a task for IberLEF 2023 composed of two subtasks that consists of using tweets multimodal information, text and images, in order to challenge participants to develop multimodal methods able to classify tweets as reporting a violent event or not (subtask 1, binary classification) and recognize the crime category (subtask 2, multiclass classification) [6] among three classes (accident, murder, and theft) for violent events reports and one class (other) for tweets of different nature.

The provided datasets were the following: Train, Validation and Test, being the last two used for development and the final stage. The train dataset consists of 2996 labeled tweets with 4267 images associated to them, the validation dataset with 582 tweets and 812 images, and the test dataset consists of 1153 tweets with 1621 images.

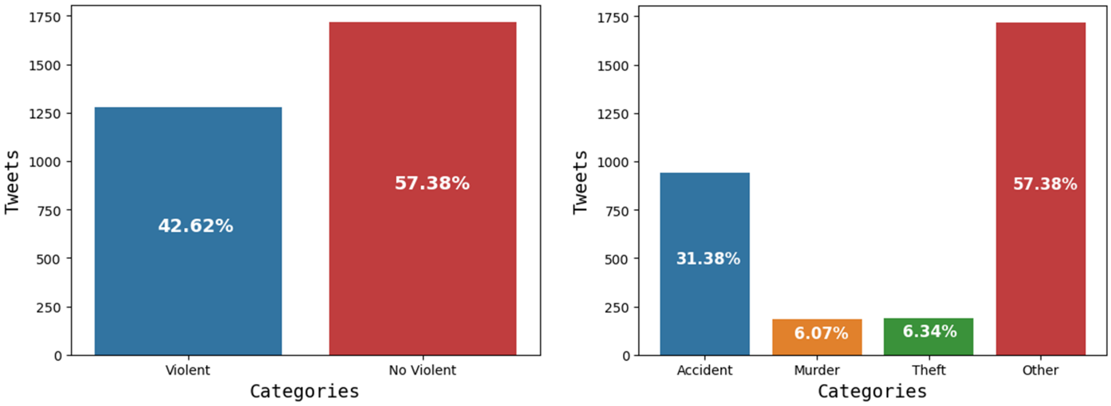
Figure 1 shows the data distribution of the different categories for the tweets presented in the training dataset. It can be seen an important data imbalanced and it is worth mentioning that from all the data only 1.13% are multi-label instances.

Detailed information about the shared task (e.g., related work, the evaluation framework, or the results of other participants) can be found in the organizers overview article [6].

3. Proposed Approach

Our participation in the DA-VINCIS shared task consisted of several experiments that included the use of only text (unimodal), and combining both types of data (multimodal).

Our main approach consisted of two key points: data augmentation using LLM for text and image captioning to extract the visual information. To retrieve new data from LLM like GPT-3



(a) Subtask 1: Binary Classification

(b) Subtask 2: Multiclass Classification

Figure 1: Data distribution for the training set.

there are two options:(i) fine tune the model with a paired list of prompts and ideal responses; and (ii) by prompt engineering. Due to the imbalanced nature of the data, we generated the synthetic tweets through prompt engineering, which consists of designing, optimizing, and refining prompts used to communicate with the Artificial Intelligence (AI) language models. Further details will be present below.

3.1. Overview of the Approach

We decided to experiment with different strategies to handle the two types of data that these tasks offered, and they went from using only text to a combination of text and images; however, we decided to only train models with data in the textual domain. This was possible by getting the captions of images associated to each tweet, as it has been explored in others works classifying images related to violent events [7]. The general workflow for this multimodal approach is seen in Figure 2, which consists of the following steps: image captioning, data augmentation, and text processing before feeding them to a BERT model.

3.2. Image Captioning

In order to get a description of the images the model BLIP was used, a Vision-Language Pre-training model pre-trained on COCO dataset-base architecture with Vision Transformer (ViT) [8] large backbone. BLIP effectively utilizes the noisy web data by bootstrapping the captions, where a captioner generates synthetic captions and a filter removes the noisy ones [9]. The use of this model allowed us to have more accurate descriptions when images were related to violent events and general topics.

An example of the captions that can be generated by BLIP using some images from the training dataset and another one from the web can be seen in Figure 3.

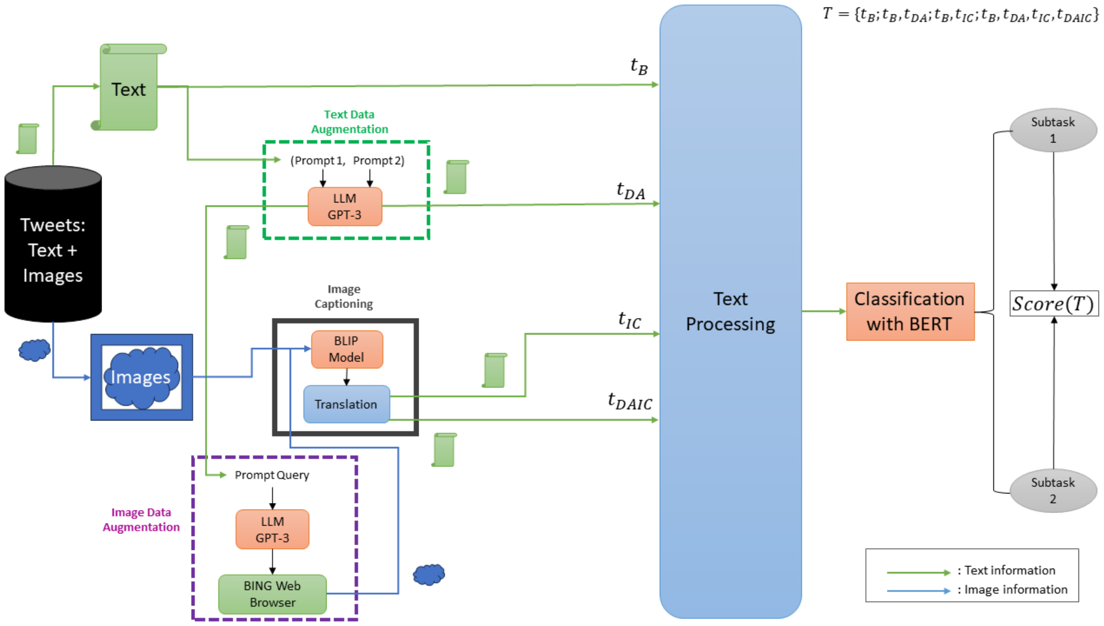
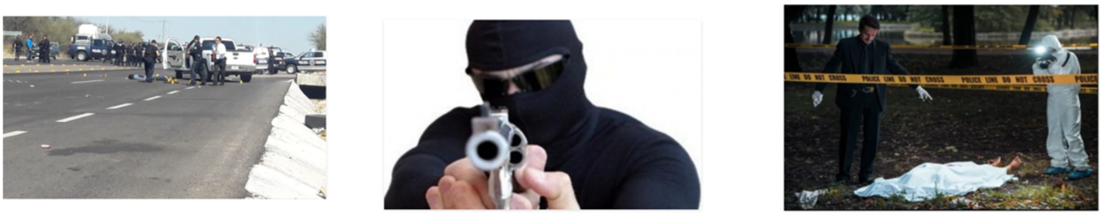


Figure 2: Workflow followed during experimentation. Where: T are all the combinations, t_B are the texts from the original tweets, t_{DA} are the texts generated by GPT-3, t_{IC} are the texts obtained from the image captioning process, and t_{DAIC} are the texts obtained from the image captioning process using the images retrieved from the data augmentation step.



"There are several police officers standing on the side of the road"

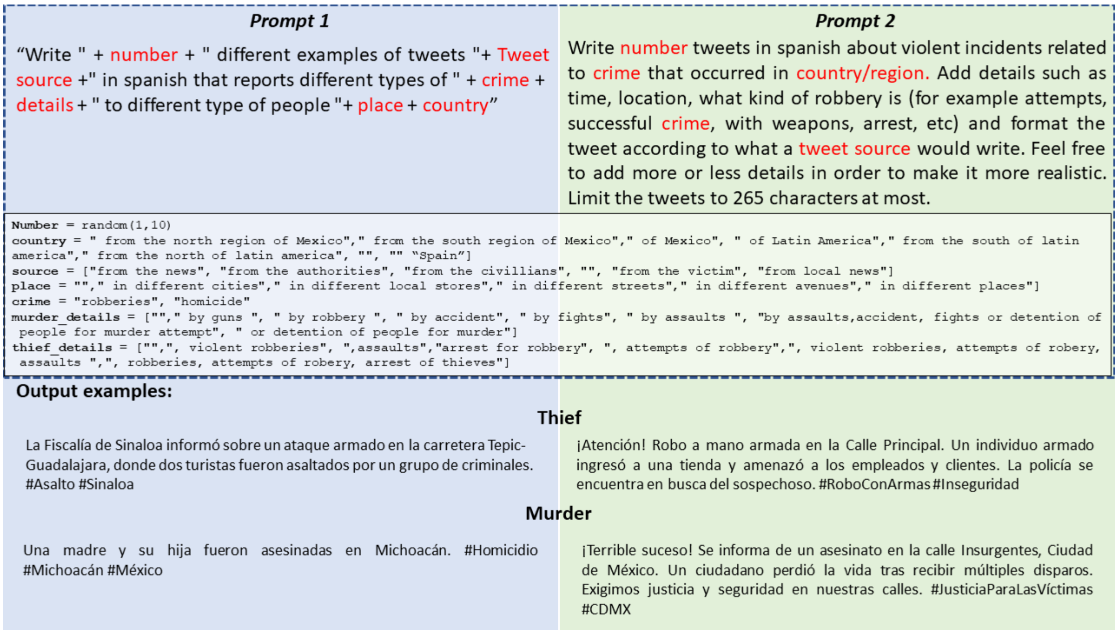
"arafed man in black mask holding a gun"

"A man in a suit and a woman in a White dress stand near a yellow tape"

Figure 3: Example of captions generated by the model BLIP.

3.3. Data Augmentation

We decided to apply data augmentation in order to handle the imbalanced data presented between classes, under the hypothesis that more data will bring a better generalization and performance for the model. In Figure 2 can be seen the general process to make data augmentation for text and images. It can be seen that image data augmentation depends on the text data augmentation step.



Prompt 1

"Write " + number + " different examples of tweets "+ Tweet source +" in spanish that reports different types of " + crime + details + " to different type of people "+ place + country"

Prompt 2

Write number tweets in spanish about violent incidents related to crime that occurred in country/region. Add details such as time, location, what kind of robbery is (for example attempts, successful crime, with weapons, arrest, etc) and format the tweet according to what a tweet source would write. Feel free to add more or less details in order to make it more realistic. Limit the tweets to 265 characters at most.

```
Number = random(1,10)
country = " from the north region of Mexico"," from the south region of Mexico"," of Mexico", " of Latin America"," from the south of latin
america"," from the north of latin america", "", "" "Spain"]
source = ["from the news", "from the authorities", "from the civillians", "", "from the victim", "from local news"]
place = ""," in different cities"," in different local stores"," in different streets"," in different avenues"," in different places"]
crime = "robberies", "homicide"
murder_details = [""," by guns ", " by robbery ", " by accident", " by fights", " by assaults ", "by assaults,accident, fights or detention of
 people for murder attempt", " or detention of people for murder"]
thief_details = ["",", violent robberies", ",assaults","arrest for robbery", ", attempts of robbery",", violent robberies, attempts of robery,
 assaults ",", robberies, attempts of robery, arrest of thieves"]
```

Output examples:

Thief

La Fiscalía de Sinaloa informó sobre un ataque armado en la carretera Tepic-Guadalajara, donde dos turistas fueron asaltados por un grupo de criminales. #Asalto #Sinaloa

¡Atención! Robo a mano armada en la Calle Principal. Un individuo armado ingresó a una tienda y amenazó a los empleados y clientes. La policía se encuentra en busca del sospechoso. #RoboConArmas #Inseguridad

Murder

Una madre y su hija fueron asesinadas en Michoacán. #Homicidio #Michoacán #México

¡Terrible suceso! Se informa de un asesinato en la calle Insurgentes, Ciudad de México. Un ciudadano perdió la vida tras recibir múltiples disparos. Exigimos justicia y seguridad en nuestras calles. #JusticiaParaLasVíctimas #CDMX

Figure 4: Structure of prompt 1 and 2 and some examples for category murder and thief. Highlight words correspond to the parameters that were changing.

3.3.1. Text

In order to increase the number of tweets we used the OpenAI LLM GPT-3 [10] API to generate synthetic tweets through what is called *Prompt engineering*. We created two different prompts with different structure. The first one, *Prompt 1*, can be considered a dynamic prompt where some parameters such as place, crime, crime details, and tweet source varies randomly among a list of pre-defined parameters. The second one, *Prompt 2*, can be considered a more static prompt where it was sought to have more detail in the prompt structure but only varying in specific words such as country, type of crime, and tweet source, having at the end two sets of synthetic data coming from the two different prompts. We first increased from 1 up to 7 times the number of original tweets for the underrepresented classes (*Murder, Thiev*) and evaluated it only using text to see if any improvement would come from this strategy. The results from local experiments showed that the best performance was scored when duplicating the data. Both prompts and some examples retrieved from them can be seen in Figure 4.

3.3.2. Image

For the images, we decided to use the web to retrieve new images related to the new instance with the help of the Bing web browser. For this, we extract relevant information using *Prompt Query* (Appendix A), named *queries*, with the help of GPT-3 model. This queries from GPT-3 had information about the place and type of violent event (Thief or Murder). Due to a lack of time, we could only experiment assigning one image per synthetic tweet.

Figure 5 shows some examples obtained from certain synthetic tweets, the queries obtained

Tweet	Query	Image
Una víctima fue dejada herida tras un asalto a una tienda en la Ciudad de México. #Asalto #Mexico	Asalto en la Ciudad de México	
La policía de Puebla reportó un ataque armado en el centro de la ciudad, donde dos hombres armados asaltaron a varios transeúntes. #Asalto #Puebla	Ataque armado Puebla	
¡Alerta! Robo a transeúnte en la Avenida Central. Una mujer fue víctima de un robo violento por parte de un ladrón en motocicleta. Mantengamos la precaución en espacios públicos.#RoboEnVíaPública #Cuidado	Robo a transeúnte Avenida Central	
¡Atroz crimen en la tienda de conveniencia de la colonia Roma! Una mujer de mediana edad perdió la vida en un intento de robo violento. Exigimos justicia y mayor vigilancia para garantizar la seguridad de los ciudadanos. #ViolenciaEnLasCalles #México	Crimen tienda conveniencia colonia Roma México	

Figure 5: Examples of the process for image data augmentation.

for the image search, and finally the image assigned.

3.4. Classification

For the classification step, some text processing had to be done. The processing steps for the text (tweets and captions) are based on [11] and consist of lower-cased all the tweets, removing URLs and emojis, only the symbol for hashtags (#'s) and user tagging (@'s), leaving the words after the hashtag and user tagging by itself, and special characters such as exclamation marks, question marks, and commas, among others.

The captions that result from the BLIP model need other steps before going to the ones described before. First, as there are some tweets with more than one image associated, we split them as if they were independent tweets, having more than one tweet but different captions at the end. Then, the images were passed to the BLIP model to get the captions for each one. To avoid some issues with captions generated, we conditioned the output to have *"An image of..."*, to prevent some noisy captions, as can be seen in Figure 3 for the image with a man holding a gun. As the captions obtained were all in English we translated them with the help of Google translator implementation in the library "Deep translator" [12]. Finally, we concatenate all the captions that corresponded to one tweet and proceed to the text processing.

In order to prepare all this information for BERT tokenizer and sequentially for BERT classification model, we merged each tweet with its corresponding image captioning. A clear representation after text pre-process and adding the special tokens can be seen in Figure 6. We

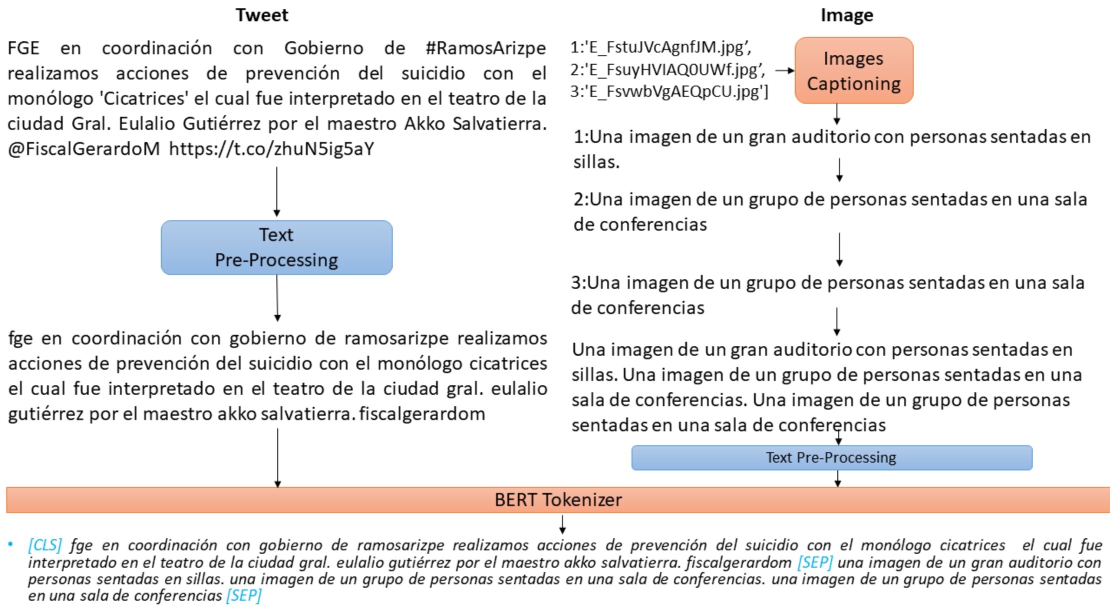

Tweet
FGE en coordinación con Gobierno de #RamosArizpe realizamos acciones de prevención del suicidio con el monólogo 'Cicatrices' el cual fue interpretado en el teatro de la ciudad Gral. Eulalio Gutiérrez por el maestro Akko Salvatierra. @FiscalGerardoM https://t.co/zhuN5ig5aY
Text Pre-Processing
fge en coordinación con gobierno de ramosarizpe realizamos acciones de prevención del suicidio con el monólogo cicatrices el cual fue interpretado en el teatro de la ciudad gral. eulalio gutiérrez por el maestro akko salvatierra. fiscalgerardom
Image
1:'E_FstuJVcAgnfJM.jpg', 2:'E_FsuyHVIAQ0UWf.jpg', 3:'E_FsvwbVgAEQpCU.jpg']
Images Captioning
1:Una imagen de un gran auditorio con personas sentadas en sillas.
2:Una imagen de un grupo de personas sentadas en una sala de conferencias
3:Una imagen de un grupo de personas sentadas en una sala de conferencias
Una imagen de un gran auditorio con personas sentadas en sillas. Una imagen de un grupo de personas sentadas en una sala de conferencias. Una imagen de un grupo de personas sentadas en una sala de conferencias
Text Pre-Processing
BERT Tokenizer
[CLS] fge en coordinación con gobierno de ramosarizpe realizamos acciones de prevención del suicidio con el monólogo cicatrices el cual fue interpretado en el teatro de la ciudad gral. eulalio gutiérrez por el maestro akko salvatierra. fiscalgerardom [SEP] una imagen de un gran auditorio con personas sentadas en sillas. una imagen de un grupo de personas sentadas en una sala de conferencias. una imagen de un grupo de personas sentadas en una sala de conferencias [SEP]

Figure 6: Example of the input for BERT model classification. From raw data to text pre-processing. The tweet has asigned 3 images.

used BERT in two ways: as two independent models per subtask and in a cascade classification method for subtask 2, further detail can be seen in Section 4.

4. Experiments Setup and Results

4.1. Data Partitioning

In order to decide which experimental setup would be better for our approach, we carried our experiments with the train dataset through a 4-folds cross validation, i.e, having 75% of the labeled data for training and 25% for testing, so in this way we could get a general idea of the performance of our model. For the development (validation dataset) and final (test dataset) phases, we worked with 100% of the training data and data configuration shown in the following sections.

4.2. Exerimental Setup

The different data configurations that we worked on and finally submitted to CodaLab are described in Table 1.

For our model, we decided to take advantage of the pre-trained transformers model BERT, more specifically BETO, as previous experiments have shown that it has a higher performance than traditional machine learning algorithms. For its hyperparameters we decided to go with some general parameters mentioned in [13]: Training the model with 3 epochs, a learning

Table 1
Submissions descriptions for subtask 1 and 2. DA# means Data Augmentation while # means either the instances from the first prompt (1), second prompt (2) or both (12).

Subtask	Strategy	Description
1, 2	Tweets	Only tweets as information input, Unimodal.
1, 2	Tweets + DA1	Tweets and synthethic data from first prompt, unimodal.
1, 2	Tweets + DA2	Tweets and synthethic data from second prompt, unimodal.
1, 2	Tweets + Captions	Tweets and captions from images, multimodal.
1, 2	Tweets + Captions + DA1 + Captions	Tweets and captions from images, with synthethic tweets from prompt 1 and captions from images from bing, multimodal.
1, 2	Tweets + Captions + DA2 + Captions	Tweets and captions from images, with synthethic tweets from prompt 2 and captions from images retrieved from bing, multimodal.
1, 2	Tweets + Captions + DA12 + Captions	Tweets and captions from images, with synthethic tweets from prompt 1 and 2 and captions from images retrieved from bing, multimodal.
2	Tweets. Cascade classification	Ensemble method using only tweets. Combination of binary classification using Tweets and captions and multiclass classification using only tweets.
2	Tweets + captions. Cascade classification	Ensemble method using tweets and captions. Combination of binary classification using Tweets and captions and multiclass classification using tweets and captions.
2	Tweets + captions + DA1 + Captions. Cascade classification	Ensemble method using tweets, captions and synthethic data from prompt 1. Combination of binary classification using Tweets and captions and multiclass classification using tweets, captions and synthethic tweets with its image captions.

rate of 4e-5, an Adam epsilon of 1e-6 and a batch size of 32. This hyperparameters stood fixed through all the process and final phase.

We decided to approach this challenge in two ways: first, with one independent model for each subtask, and second, with an ensemble cascade classification method for subtask 2. For the cascade classification, we took the best performance for subtask 1 shown in the validation dataset, which was using a combination of tweets and image captions with no data augmentation of any kind. The predictions that were classified as part of the category *violent* were moved to a second model sequentially, which was trained with only the instances categorized as violent (being trained with 3 classes), in such a way that this last model predicts which of the three the tweet belongs to, making it easier for the model in a certain way by reducing the option from four to three but carrying the mistakes from the binary classification stage.

Table 2
F1-Score for each submission for subtask 1(BIN) and 2(MULT) for the Test dataset.

Strategy	F1-Score BIN	F1 Score MULT
Tweets	0.9175	0.8575
Tweets + DA1	0.9076	0.8619
Tweets + DA2	0.9116	0.8536
Tweets + Captions	**0.9203**	**0.8797**
Tweets + Captions + DA1 + Captions	0.9187	0.8719
Tweets + Captions + DA2 + Captions	0.9151	0.8617
Tweets + Captions + DA12 + Captions	0.9127	0.8649
Tweets. Cascade classification	–	0.8741
Tweets + captions. Cascade classification	–	0.8724
Tweets + captions + DA1 + Captions. Cascade classification	–	0.8778

4.3. Results

In this section, we present our final results in the test dataset and the final strategies that resulted in the best performance in our cross validation and development phase, concluding that the best scores were the ones related to the text domain. For subtask 1, we did a total of 7 unique submissions, whereas for multiclass we submitted 10 of them.

Table 2 shows the official results for each strategy that was submitted to CodaLab. From all the strategies, we can see that the best performance for subtask 1 and subtask 2 are obtained using multimodal information, with text from tweets and text from image captions. In order to tackle the imbalanced training dataset, we hypothesized that more data would help the model generalize better. Our experiments with GPT-3 to generate synthetic tweets showed some improvements when only considering tweets, and this can be clearly seen in Table 2 for subtask 2, as the augmentation was made considering this issue, but it brings a slight dissatisfaction to the obtained result when using multimodal information. We attribute this to the data from the image data augmentation, which received less time and experimentation compared to the text due to time constraints. In Appendix B can be found some noisy images. On the other hand, the second best performance for each subtask is not far from more than a few decimals and was achieved using a multimodal strategy with data augmentation and also using a cascade classification approach for subtask 2. However, as we have no access to the other metrics of precision and recall for all our submissions, we can not determine how bad or good the data augmentation approach was.

For the official results of the binary subtask, we ranked in second place with a F1-Score of 0.9203 with BERT fine-tuned with the original data and image captions. This approach was also the best among the standalone in the development phase. We also achieved the best recall among all participants with a score of 0.9409; however, the precision for this task ranked in 8th place with a score of 0.9006.

In the case of the multiclass subtask, the submission that achieved the best performance with an F1-Score of 0.8797 was, again, the combination of tweets with image captions, scoring the best precision with 0.8737 and the 3rd best recall, a more balanced metrics compared with the precision and recall of subtask 1. In this regard, we conclude that this method tends to

Table 3
Official results for subtask 1 on the Test dataset.

Strategy	F1-Score	Precision	Recall
1st Best team	0.9264	0.9302	0.9226
Tweets + Captions	**0.9203**	**0.9006**	**0.9409**
3rd Best team	0.9186	0.9067	0.9308
Baseline	0.8948	0.9456	0.8493

Table 4
Official results for subtask 2 on the Test dataset.

Strategy	F1-Score	Precision	Recall
Tweets + Captions	**0.8797**	**0.8737**	**0.8864**
2nd Best team	0.8733	0.8523	0.8973
3rd Best team	0.8698	0.8622	0.8784
Baseline	0.8427	0.7663	0.9407

favor recall (less false negatives) over precision. For subtask 1, our results in the test dataset remains consistent with what has been seen in the validation dataset and in our cross validation; however, for subtask 2 the best performance in the validation dataset was using only text from tweets with data augmentation, followed by a cascade classification using tweets and image captions. The best results are shown in Tables 3 and 4 for each subtask.

5. Conclusions

This paper describes our participation at the DA-VINCIS IberLEF 2023 challenge on both subtasks, Violent Event Identification and Violent Event Categorization. Our participation focused more on the type of data than the technique, having the best result in the text domain using tweets and captions automatically extracted from the images. This can be explained, at least from our previous experiments perspective, because treating the visual information from the tweets as images with an image model classifier could be a challenge as some classes could not be easily distinguished as if they were sports, food, or another object for classification. This may be due to the fact that, first, the images come from social media and second, some images, for example, related to robbery, can be confused with an image related to murder in the case that the authors of the tweet presented an image post-event.

Our submitted experiments showed one of the best performances, especially for subtask 2. However, there is still a chance to improve it through hyperparameter optimization, as we kept the same hyperparameters in all of our experiments and submissions. We performed data augmentation for the two domains, text and image, with the idea of improving the model's performance. For text, on subtask 2, the improvement was apparent (see Table 2), but when considering image captions, it did not show any. This was expected, as we did more data augmentation experiments with text than with images, and some images retrieved from the web could have added some noise rather than helped the models. We also believe that the images assignment for the synthetic tweets can be improved through another way, like with image generation models like DALL-E instead of depending on the web, as noisy images can be assigned with our method. At the same time, as with every other data augmentation technique, using GPT-3 as a data augmentation does not ensure an improvement just by using it, it requires some time on prompt engineering to get data that fits your needs but as far as our experiments in this topic go, increasing too much the data will not necessarily bring better performance, as we got our best behavior augmenting in the range of two to four times more the base number of the minority instances.

For future work, we plan to try our prompts with recent versions of LLM, such as GPT-4 or even BART from Google for text data augmentation, as well as improve them. For the visual information, we plan to test other techniques to retrieve relevant information from tweets to be used to either generate images with text for visual models like Midjourney or Dall-E, or to even explore more of the potential of the web by incorporating other strategies such as more images or filters to reduce the amount of noisy images retrieved. In terms of models, explore in greater depth the Vision Transformer or other model to process visual information to see the capacities and limitations when considering two different types of data at the same time and comparing them when only working in one domain.

Acknowledgments

This work was supported by National Council of Humanities Sciences and Technologies of Mexico (CONAHCYT) under the grant 805444.

References

[1] A. M. Kaplan, M. Haenlein, Users of the world, unite! The challenges and opportunities of Social Media, Business Horizons 53 (2010) 59–68. doi:10.1016/j.bushor.2009.09.003.

[2] R. Prieto Curiel, S. Cresci, C. I. Muntean, S. R. Bishop, Crime and its fear in social media, Palgrave Communications 6 (2020) 1–12. URL: http://dx.doi.org/10.1057/s41599-020-0430-7. doi:10.1057/s41599-020-0430-7.

[3] F. Mata, M. Torres-Ruiz, G. Guzman, R. Quintero, R. Zagal-Flores, M. Moreno-Ibarra, E. Loza, A Mobile Information System Based on Crowd-Sensed and Official Crime Data for Finding Safe Routes: A Case Study of Mexico City, Mobile Information Systems 2016 (2016). doi:10.1155/2016/8068209.

[4] S. P. Sandagiri, B. T. Kumara, B. Kuhaneswaran, Deep Neural Network-Based Approach to Identify the Crime Related Twitter Posts, 2020 International Conference on Decision Aid Sciences and Application, DASA 2020 (2020) 1000–1004. doi:10.1109/DASA51403.2020.9317098.

[5] C. A. Piña-García, L. Ramírez-Ramírez, Exploring crime patterns in Mexico City, Journal of Big Data 6 (2019). URL: https://doi.org/10.1186/s40537-019-0228-x. doi:10.1186/s40537-019-0228-x.

[6] H. Jarquín-Vásquez, D. I. Hernández-Farías, J. Arellano, H. J. Escalante, L. Villaseñor-Pineda, M. Montes-y-Gómez, F. Sanchez-Vega, Overview of DA-VINCIS at IberLEF 2023: Detection of Aggressive and Violent Incidents from Social Media in Spanish, Procesamiento del Lenguaje Natural 71 (2023).

[7] O. Arriaga, P. Plöger, M. Valdenegro-Toro, Image Captioning and Classification of Dangerous Situations (2017). URL: http://arxiv.org/abs/1711.02578. arXiv:1711.02578.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2020). URL: http://arxiv.org/abs/2010.11929. arXiv:2010.11929.

[9] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (2022). URL: http://arxiv.org/abs/2201.12086. arXiv:2201.12086.

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners (2020). arXiv:2005.14165.

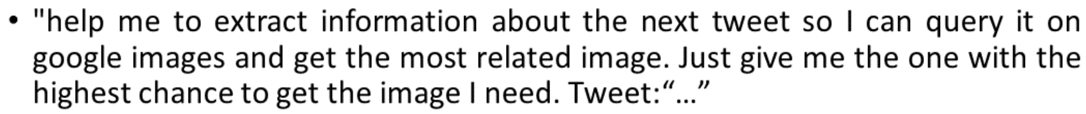
[11] U. Naseem, I. Razzak, P. Eklund, A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter, Multimedia Tools and Applications 80 (2021) 1–28. doi:10.1007/s11042-020-10082-6.

[12] M. LLC, Deep translator, 2020. URL: https://deep-translator.readthedocs.io/en/latest/.

[13] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies - Proceedings of the Conference 1 (2019) 4171–4186. arXiv:1810.04805.

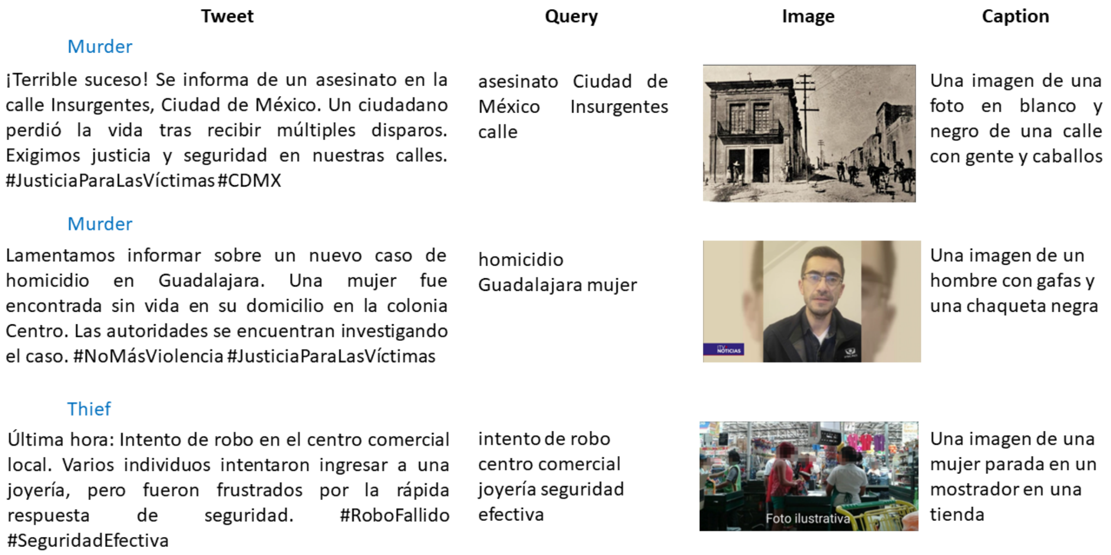
A. Prompt Query



- "help me to extract information about the next tweet so I can query it on google images and get the most related image. Just give me the one with the highest chance to get the image I need. Tweet:"…"

Figure 7: Prompt used to extract information from synthethic tweets.

B. Noisy Images



Tweet	Query	Image	Caption
Murder ¡Terrible suceso! Se informa de un asesinato en la calle Insurgentes, Ciudad de México. Un ciudadano perdió la vida tras recibir múltiples disparos. Exigimos justicia y seguridad en nuestras calles. #JusticiaParaLasVíctimas #CDMX	asesinato Ciudad de México Insurgentes calle		Una imagen de una foto en blanco y negro de una calle con gente y caballos
Murder Lamentamos informar sobre un nuevo caso de homicidio en Guadalajara. Una mujer fue encontrada sin vida en su domicilio en la colonia Centro. Las autoridades se encuentran investigando el caso. #NoMásViolencia #JusticiaParaLasVíctimas	homicidio Guadalajara mujer		Una imagen de un hombre con gafas y una chaqueta negra
Thief Última hora: Intento de robo en el centro comercial local. Varios individuos intentaron ingresar a una joyería, pero fueron frustrados por la rápida respuesta de seguridad. #RoboFallido #SeguridadEfectiva	intento de robo centro comercial joyería seguridad efectiva		Una imagen de una mujer parada en un mostrador en una tienda

Figure 8: Example of some noisy images retrieved from the web.