Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 351-360 recibido 30-06-2023 revisado 08-07-2023 aceptado 17-07-2023

Overview of DA-VINCIS at IberLEF 2023: Detection of Aggressive and Violent Incidents from Social Media in Spanish

Resumen de la Tarea DA-VINCIS en IberLEF 2023: Detección de Incidentes Violentos en Redes Sociales en Español

Horacio Jarquín-Vásquez,[1] **Delia Irazú Hernández-Farías,**[1]
Luis Joaquín Arellano,[1] **Hugo Jair Escalante,**[1] **Luis Villaseñor-Pineda,**[1,2]
Manuel Montes-y-Gómez,[1] **Fernando Sanchez-Vega**[3,4,5]
[1]Laboratorio de Tecnologías del Lenguaje (INAOE), Mexico
[2]Centre de Recherche GRAMMATICA (EA 4521), Université d'Artois, France
[3]Mathematics Research Center (CIMAT), Guanajuato, Mexico
[4]El Colegio de México (COLMEX), Mexico
[5]Consejo Nacional de Ciencia y Tecnología (CONACYT), Mexico.
{horacio.jarquin, dirazuhf, arellano.luis, hugojair, villasen, mmontesg}@inaoep.mx
fernando.sanchez@cimat.mx

Abstract: In this paper, we present the overview of the DA-VINCIS 2023 shared task which was organized at IberLEF 2023 and co-located in the framework of the 39^{th} International Conference of the Spanish Society for Natural Language Processing (SEPLN 2023). The main aim of this task is to promote the research on developing automatic solutions for detecting violent events in social networks. Two subtasks were considered: (i) A binary classification task aimed to determine whether or not a tweet is about a violent incident; and (ii) A multi-label multi-class classification task in which the category(ies) of a violent incident must be identified. A multimodal manual annotated corpus comprising both tweets and images associated to them was provided to the participants. A total of 15 systems were submitted for the final evaluation phase. Competitive results were obtained for both subtasks, the higher ones were in the binary classification task. Corpora and results are available at the shared task website at `https://codalab.lisn.upsaclay.fr/competitions/11312`.
Keywords: DA-VINCIS, violent event detection, text classification.

Resumen: En este documento presentamos el resumen de la tarea DA-VINCIS 2023 organizada como parte del IberLEF 2023 junto con la 39^{a} Conferencia Internacional de la Sociedad Española para el Procesamiento de Lenguaje Natural (SEPLN 2023). El principal objetivo de la tarea es promover la investigación en el desarrollo de soluciones automáticas para la detección de incidentes violentos en redes sociales. Se consideraron dos tareas: (i) Clasificación binaria cuyo objetivo es determinar si un tweet reporta o no un incidente violento, y (ii) Clasificación multietiqueta multiclase donde la categoría(s) de un indicente violento debe ser identificada. A los participantes de la tarea se les proporcionó un conjunto de datos multimodal anotado manualmente el cual contiene tanto tweets como imágenes relacionadas. Un total de 15 sistemas fueron enviados para la fase de evaluación. En ambas tareas se obtuvieron resultados competitivos, siendo los mejores aquellos de la tarea de clasificación binaria. El corpus y los resultados detallados pueden consultarse en el sitio web de la tarea: `https://codalab.lisn.upsaclay.fr/competitions/11312`.
Palabras clave: DA-VINCIS, detección de eventos violentos, clasificación de textos.

ISSN 1135-5948 DOI 10.26342/2023-71-27

1 Introduction

Violent incidents are among the factors that can undermine the well-being of society. Violence provokes negative effects on people who directly suffer from it and also in those who witness it (such as for example in their mental health). Governments are in charge of ensuring the safety and security of society, therefore, they must procure to have different alternatives to deal with violent incidents. Nowadays, social media are one of the main communication and information channels. On these platforms a wide range of data is generated, shared, and thus consumed by people in real time. Such content represents a powerful information source for permanently detecting and monitoring events happening. Therefore, at any time violent incidents are very likely to be present in user-generated content.

Computational linguistics approaches could be exploited as a way to develop automatic tools for timely detecting violent incidents in social media. This alternative could serve different purposes, for example: a) Improving the response rate by the authorities when a violent event happens and it is posted on social media; b) Generating policies aiming to prevent violent incidents taking into account how and where these events occur; and, c) Keeping the social media users informed about violent incidents occurring in their surroundings. In spite of being a very important topic, only a few efforts have been done in this direction.

Twitter data have been already used for investigating on this topic. An approach combining data from Twitter and official statistics was proposed with the aim of providing safe routes to increase the confidence levels of citizens in Mexico City (Mata Rivera et al., 2016). An analysis to assess the usefulness of using Twitter data to discover the spatial distribution of crime frequency in Mexico City was presented in (Piña-García and Ramírez-Ramírez, 2019). Classifying tweets according to a given crime type by means of a method based on neural networks and active learning was proposed (Sandagiri, Kumara, and Kuhaneswaran, 2020). For retrieving Twitter data, some keywords like guns, fight, robbery, kidnapping, thief, etc. were used.

Last year, the first edition of the DA-VINCIS shared task was organized (Arellano et al., 2022). Two challenges were proposed in the framework of this task: i) to determine whether or not a tweet describes a violent incident, and ii) to identify which kind of event is expressed (if any). Participants were provided with a corpus of tweets accordingly labeled. A total of 12 teams participated in the shared task, with a wide range of approaches like pre-trained transformers, ensembles, multi-task learning, advanced linguistic features, and prompt learning, among others. Besides, for tackling data scarcity, data augmentation methods were applied. According to the official ranking, better results were obtained for the first subtask in comparison with the last one.

This year, we organized the second edition of the DA-VINCIS shared task collocated in the framework of the IberLEF2023. DA-VINCIS 2023 is aimed at the detection of violent incidents on Twitter, comprising the same subtasks as in the previous edition. However, instead of only providing textual data the participants were provided with a multimodal dataset consisting of tweets associated with at least an image. As in the previous edition, data provided to the participants is written in Mexican Spanish. In addition, we are also interested in promoting research in detecting violent events which is a very relevant topic that can have a great impact on the whole of society.

The remainder of this paper is organized as follows. Section 2 describes the dataset developed and related details of the shared task. Section 3 presents a summary of the approaches proposed by the participating systems as well as the obtained results. Finally, our conclusions and future work are exposed in Section 4.

2 Task description

2.1 Dataset

The DA-VINCIS 2023 corpus is an upgrade of the dataset used in the previous edition (Arellano et al., 2022). It is composed of Twitter data associated with reports of violent incidents in Mexican Spanish. The main difference in this renewed version lies in the kind of data considered: all tweets in this corpus have at least an image associated. A second difference is on the categories considered as well as on the manual annotation performed. First, the *Kidnapping* category was excluded due to the low number of instances of this category in the first version of the dataset which provokes a very skewed class imbalance distribution. On the other hand, the former category *Homicide*, which was considered as the act of "Deprivation of life" was redefined to *Murder* by adding the condition of the intentional purpose on it leaving out situations like cause death in an accident and so on.

Furthermore, this year we decided to pay special attention to a crucial aspect for timely dealing with violent incidents: the time frame elapsed between the occurrence of the event and when posted on social media, in this case, a lapse of 24 hours was defined as the maximum time for considering a tweet as reporting this kind of events. It is worth mentioning that in the dataset of the previous edition, any restriction on the time frame was established.

In the corpus belonging to the DA-VINCIS 2023 edition the following four categories The following categories were considered in the InDA-VINCIS 2023 corpus:

- *Accident.* An unexpected event or action which results in involuntary damage to people or their environment.
- *Murder.* The act of depriving the life of a person intentionally.
- *Robbery.* The event of seizing or willfully destroying the properties of someone else without the right or the consent of the person who can legally dispose of them by using force or threat.
- *Other.* Other kinds of violent incidents different from the above as well as content non-related to violent incidents.

Manual Annotation

Once defined the new set of categories, a manual annotation process was performed. DA-VINCIS 2022 dataset was used as starting point. From it, those instances labeled by the less reliable annotators were discarded. For determining which of the annotators are considered as the less reliable, their annotations were compared against a gold standard annotated by the organizers of the tasks, those showing the highest differences with respect to this subset are considered as the less reliable. The remaining instances were fully relabeled. We asked a group of annotators to carefully read each tweet and to assign a corresponding label. For doing this task, we provide guidelines including the aforementioned definitions of each category and also to make emphasis considering the 24-hour timeframe. It is important to mention that, assigning a tweet with more than one label was allowed except in the case of *Other* with cannot be combined with the remaining labels. Aiming to avoid the annotators being influenced by the images of the tweets (as it seems to have happened last year) during the labeling process the annotators only had access to the text.

Around 5,000 tweets were manually annotated during the first round, those instances where the annotators disagreed were passed through a second round of labeling performed by another group of annotators. In the end, for subtask 1 we have a total of 2005 tweets labeled as reporting a violent incident (denoted as *positive*) and 2726 of the *Other* class, from these instances, a total of 2996 instances were allocated to the training set, 582 instances were assigned to the validation set, and the remaining 1153 instances were designated for the test partition. For what concerns subtask 2, Table 1 shows the distribution of the tweets across the different categories. Furthermore, in the last column of Table 1, we include the inter-annotator agreement in terms of the Kappa coefficient in each of the kinds of violent incidents. As can be noticed, the highest agreement was reached in the *Accident* class probably because this kind of incident is the easiest to identify. On the other hand, *Murder* is the one where annotators disagree most, maybe due to the different aspects involved to recognize an intentional act.

	Train	Test	Total	*kappa*
Accident	1111	351	1462	0.92
Murder	221	82	303	0.67
Robbery	220	67	287	0.78
Other	2064	662	2726	

Table 1: Data distribution of train and test partitions. Last column shows the *kappa* value obtained during manual annotation.

As mentioned before, in the 2023 edition of the dataset, all tweets are associated with at least one image. Figure 1 shows the distribution of how many images each instance in the dataset has. As it can be observed, most tweets have only one associated image. The maximum number of images associated with a single tweet is 4. This aspect is important since the systems could take advantage (when possible) of more than one image to perform the task at hand. Tables 6, 7, and 8 showcase some examples of these tweets alongside their corresponding images.

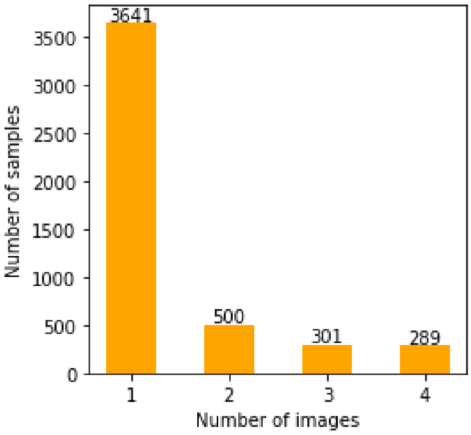


Figure 1: Distribution of how many tweets have one, two, three, or four different images associated to them.

2.2 Subtasks

DA-VINCIS 2023 encompasses two tracks: i) A *binary* classification subtask, in which the participants were asked to determine whether or not a given tweet is reporting a violent incident, and ii) A *multi-class multi-label* classification task, where the aim is to determine the kind (s) of violent incident being reported in tweets.

Both subtasks were evaluated using the DA-VINCIS 2023 corpus. The CodaLab platform (Pavao et al., 2022) was used to run the challenge. The shared task was divided into the following two phases:

- **Development phase.** Both labeled training data and unlabeled validation data were available to participants. Participants were able to submit to the CodaLab website their predictions for the validation set during this phase receiving a quickly evaluation.

- **Final phase.** Unlabeled test results were made available to participants. During the contest, they could upload up to ten submissions. Participants were ranked based on their performance on the test set.

For evaluation purposes, we consider the recall, precision, and f_1 score with respect to the *violent-incident* class as evaluation measures for subtask 1. Concerning subtask 2, the macro average recall, precision, and f_1 score were considered. Being the latter the score serving as the leading evaluation measure in both subtasks.

2.3 Baselines

As baselines methods three established approaches known for their strong performance in tasks involving image and text classification were chosen. The first baseline approach focuses solely on the text modality and involves the fine-tuning of the pre-trained BETO[1] model. Using only the visual modality a second baseline was defined. It entails fine-tuning the pre-trained Vision Transformer[2] (ViT) model. Since the DA-VINCIS 2023 dataset includes instances where more than one image is associated to each tweet, we decided to use only the first image for establishing the baseline. Lastly, our third baseline method integrates both modalities (visual and textual) by employing an early fusion technique that concatenates the classification vectors obtained from BETO and ViT models. Once more, only one image was associated with the tweets despite how many of them were available. For the sake of readability, the obtained results of these baselines were included and are referred to as Baseline (TXT), Baseline (IMG), and Baseline (TXT + IMG).

3 *Overview of Participating Systems*

The subsequent subsections provide an overview of the primary concepts explored by the participating systems, followed by an overall evaluation of their findings.

3.1 Systems' Descriptions

A total of 7 teams participated in the DA-VINCIS shared task and submitted a working note describing their solution. Since this shared task involved utilizing both text and images to identify aggressive and violent events, two of the participating teams exploited these modalities in their approaches,

[1] `https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased`

[2] `https://huggingface.co/docs/transformers/model_doc/vit`

while the remaining teams relied solely on textual information. The top-performing teams in each subtask exclusively concentrated on their respective subtask. This suggests that each subtask presents distinct challenges, indicating the inadvisability of adopting a uniform strategy for both subtasks.

Various text preprocessing techniques have been employed across both subtasks. Certain teams opted to convert all text to lowercase letters, as well as eliminate symbols, special characters, and URLs from the text (Ponce-León and López-Nava, 2023; Graff et al., 2023; Cabada et al., 2023). In order to remove out of the vocabulary words/symbols, other teams focused on removing emojis and hashtags (Ponce-León and López-Nava, 2023; Gutiérrez-Megías et al., 2023). Conversely, some teams chose to normalize user mentions and URLs, while converting emojis to textual representations (Gutiérrez-Megías et al., 2023; Cabada et al., 2023), aiming to extract the maximum amount of textual information from each tweet. Additionally, several participants applied tokenization and lemmatization techniques, as well as the removal of empty words (Hernández-Minutti et al., 2023; Rubio, Almeida, and Segura-Bedmar, 2023).

Approach	CICESE	CIMAT	ESCOM	INGEOTEC	ITC	SINAI	UC3M	csuazob
Transformers	✓	✓			✓	✓	✓	✓
BoW, n-grams, and TML			✓	✓				✓

Table 2: General approach of teams. TML stands for Traditional Machine Learning.

Table 2 provides a summary of the overall approaches employed by the participating teams. These approaches can be categorized into two main categories: transformer-based approaches and Traditional Machine Learning (TML) approaches utilizing Bag-of-Words (BoW) and n-gram representations. As observed, six out of the eight teams relied on transformer-based approaches, leveraging their exceptional performance across various Natural Language Processing Tasks (Lin et al., 2022).

Regarding transformer-based approaches, the CICESE team (Ponce-León and López-Nava, 2023) utilized fine-tuned BETO models for both the binary and multiclass subtasks. They extensively explored data augmentation techniques, including leveraging GPT-3 for the data augmentation of the text modality and conducting web searches to obtain related images. Additionally, they employed image captioning using the BLIP model to provide the model with additional contextual information. In a similar vein, the CIMAT team (Vallejo-Aldana, López-Monroy, and Villatoro-Tello, 2023) employed pre-trained models and employed BLIP for image captioning. They combined captions and the original tweet text using a separator and passed them through the RoBERTa Transformer model. As for the ITC team (Cabada et al., 2023), their approach involved the use of two Transformer-based models: BERT for Subtask 1 and RoBERTa for Subtask 2.

The SINAI team (Gutiérrez-Megías et al., 2023) employed RoBERTa Large to represent text and the BEIT Base patch16-244 model for image captioning. They also applied various data augmentation techniques, including back-translation and image modifications. Specifically, they adapted an English model to the Spanish task. The UC3M team (Rubio, Almeida, and Segura-Bedmar, 2023) investigated diverse strategies for combining Transformer language features and embeddings, utilizing the BETO tokenizer to ensure effective text tokenization. Furthermore, the csuazob team employed a frozen XML-RoBERTa model alongside logistic regression, incorporating transfer learning with a fine-tuned RoBERTa model for sentiment analysis. They also made use of back-translation techniques. Overall, the teams' employed techniques such as fine-tuning, data augmentation, image captioning, fusion techniques, and transfer learning to effectively address the identification of aggressive and violent incidents on social media.

Concerning the TML approaches, The ESCOM team (Hernández-Minutti et al., 2023) utilized various machine learning methods, including Logistic Regression, Support Vector Machines, Naive Bayes, Multilayer Perceptron, and XGBoost. They further proposed an Ensemble voting schema to enhance the performance of the individual classifiers. On the other hand, the INGEOTEC team (Graff et al., 2023) employed a stack generalization approach using the EvoMSA framework (Graff et al., 2020), which unifies the predictions of four base classifiers.

These classifiers included two BoW models and two dense BoW models with different token selection procedures. Additionally, the dense BoW representation with normalized frequency incorporated customized keywords.

3.2 Evaluation campaign results

Table 3 presents the obtained results by the participating teams in Subtask 1, which involves the binary identification of violent incidents. The teams are ranked in descending of f_1-score for the positive class (i.e., the violent incident class). Precision and Recall values are also provided to facilitate a comprehensive interpretation of these findings. Notably, the results highlighted in gray correspond to the teams that submitted working notes containing descriptions of their systems. Additionally, we have included the results obtained from our three baseline approaches.[3]

Subtask 1: Binary violent event identification

Team	Precision	Recall	F1-Score
CIMAT	**0.9302**	0.9226	**0.9264**
CICESE	0.9006	**0.9409**	0.9203
UC3M	0.9067	0.9308	0.9186
Baseline (TXT + IMG)	0.9220	0.9145	0.9182
SINAI	0.8951	0.9389	0.9165
csuazob	0.8939	0.9267	0.9100
Baseline (TXT)	0.8767	0.9409	0.9077
Arnold	0.9014	0.9124	0.9069
rkcd	0.9000	0.8982	0.8991
ITC	0.9081	0.8859	0.8969
INGEOTEC	0.9053	0.8758	0.8903
ESCOM	0.8952	0.8697	0.8822
BrauuHdzm	0.8952	0.8697	0.8822
Thisjesusalan	0.8649	0.8737	0.8693
d121201	0.9183	0.7556	0.8290
PabloGP	0.8782	0.7780	0.8251
Baseline (IMG)	0.8114	0.7800	0.7954
pakapro	0.4398	0.4908	0.4639

Table 3: Results of the participant teams in Subtasks 1. Results in bold correspond to the best results of each measure.

The CIMAT team achieved the highest performance in Subtask 1 (Vallejo-Aldana, López-Monroy, and Villatoro-Tello, 2023), followed by CICESE (Ponce-León and López-Nava, 2023). These two approaches share the commonality of utilizing both modalities, demonstrating the advantage of leveraging the complementarity provided by textual and visual information. Moreover, both approaches employed data augmentation techniques, pre-trained transformer models, and image captioning using the BLIP model as supplementary information for their systems. The UC3M team (Rubio, Almeida, and Segura-Bedmar, 2023) ranked at the third-best position, and similarly to the CICESE team, both approaches focused on a fine-tuned BETO-based model. In terms of overall performance among the various approaches, it is evident that the Transformer-based methods outperformed traditional machine learning approaches that relied on BoW representations.

It is important to highlight that the various approaches yielded remarkably similar results, with the highest performance being only 5.01% greater than the lowest performance achieved. Furthermore, the disparity between the F1-scores of the first and second-place teams is only 0.0061. In order to further analyze the differences in performance we performed an statistical analysis using the tool by (Nava-Muñoz, Graff-Guerrero, and Escalante, 2023). Figure 2 shows confidence intervals to the mean with 95% confidence for the f_1, precision and recall measures, 1000 bootstrap samples were considered (details can be found in (Nava-Muñoz, Graff-Guerrero, and Escalante, 2023)). From this figure it can be confirmed that differences between the top-4 ranked teams in terms of f_1 measure are not statistically significant.

The obtained results in Subtask 2, i.e., the violent event category recognition are shown in Table 4. The highest performance in this subtask is attributed to the CICESE team, closely followed by the SINAI team (Gutiérrez-Megías et al., 2023). These teams primarily focused on employing pre-trained transformer models and implementing data augmentation techniques in both image and text modalities. The UC3M team secured the third-best performance, relying on a pre-trained BETO-based model for their solution. Notably, the CICESE team also obtained the second position in Subtask 1, indicating the effectiveness and robustness of their approach to the task of identifying aggressive and violent incidents in social media.

Figure 3 shows confidence intervals for the violence categories, the behavior is similar as for the binary case, so it cannot be concluded that there are statistically significant differences among the top-4 teams.

[3] In this paper, we report the median performance over 5 runs for each baseline, as it offers a more reliable estimation of their performance.

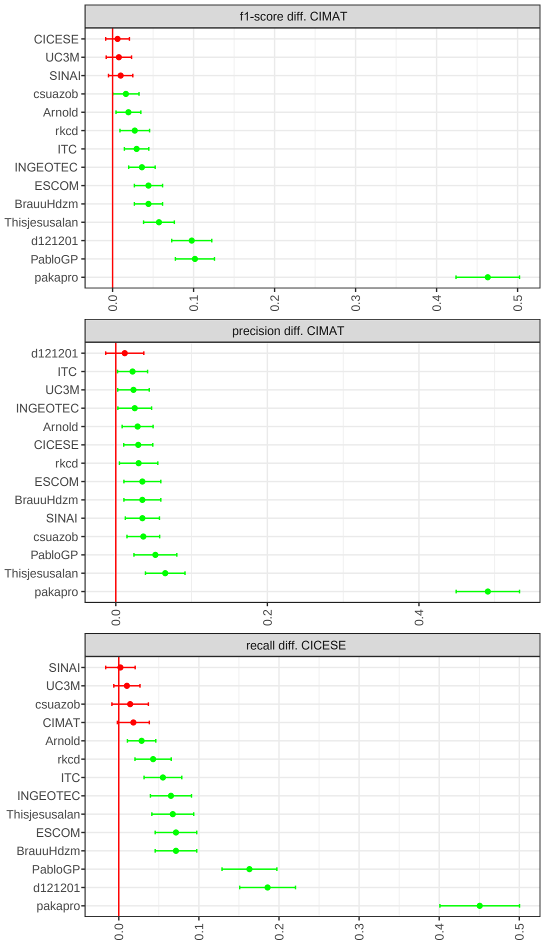


Figure 2: 95% confidence intervals for subtask 1 using bootstrapping.

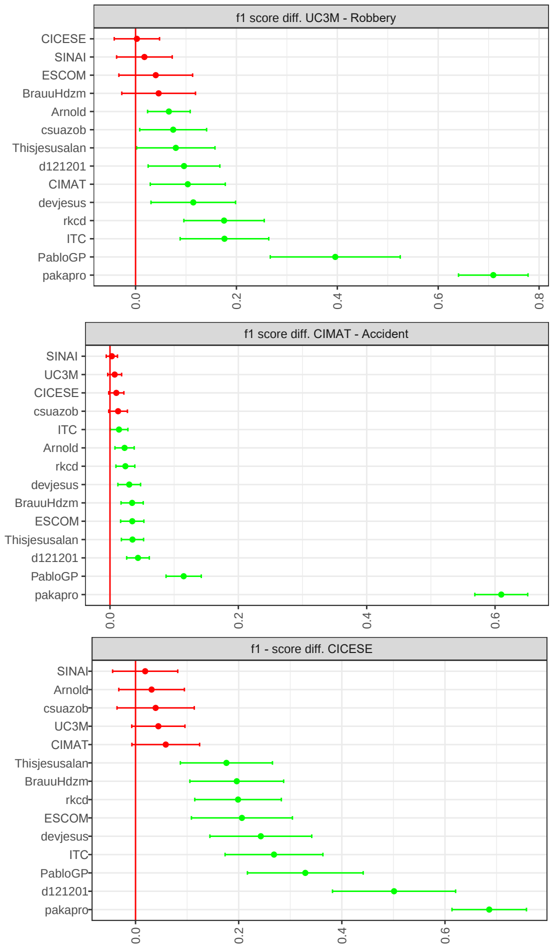


Figure 3: 95% confidence intervals for subtask 2 using bootstrapping.

3.3 Analysis

The analysis of the results reveals a notorious disparity in terms of difficulty between Subtask 2 and Subtask 1, which aligns with expectations arising from the data-skewed imbalance observed within certain categories. This discrepancy is evident in the substantial disparity between the highest and lowest reported results, with the former surpassing the latter by a notable margin of 15.03%.

In order to conduct a more comprehensive analysis of the obtained results by the participants, we first focused on examining the complementarity and diversity observed in their predictions. To quantify the level of complementarity, we employed the *Maximum Possible Accuracy (MPA)* metric, which is defined as the ratio of correctly classified instances to the total number of test instances. In this case, an instance is considered correctly classified if at least one of the participating teams correctly classified it. For measuring diversity, we utilized the *Coincident Failure Diversity (CFD)* metric (Tang, Suganthan, and Yao, 2006), which focuses on calculating the

Subtask 2: Violent event category recognition			
Team	Precision	Recall	F1-Score
CICESE	**0.8737**	0.8864	**0.8797**
SINAI	0.8523	**0.8973**	0.8733
UC3M	0.8622	0.8784	0.8698
Baseline (TXT + IMG)	0.8168	0.8991	0.8548
Arnold	0.8305	0.8715	0.8492
csuazob	0.8441	0.8577	0.8490
CIMAT	0.8394	0.8449	0.8421
Baseline (TXT)	0.7663	0.9195	0.8317
BrauuHdzm	0.8027	0.8091	0.8048
ESCOM	0.8178	0.7974	0.8036
Thisjesusalan	0.7802	0.8294	0.8030
devjesus	0.8184	0.7517	0.7789
rkcd	0.7812	0.7781	0.7773
ITC	0.7760	0.7571	0.7647
d121201	0.7306	0.7210	0.7116
PabloGP	0.7338	0.6198	0.6581
Baseline (IMG)	0.4988	0.6155	0.5495
pakapro	0.2531	0.4992	0.2860

Table 4: Results of the participant teams in Subtask 2. Results in bold correspond to the best results of each measure.

error diversity among the participant's predictions. The minimum value of this metric is 0, indicating that all teams simultaneously predict a pattern correctly or wrongly, while the maximum value is 1, representing unique misclassifications.

The results of applying the MPA and CFD metrics to assess the performance of all participating teams in the task of identifying aggressive and violent incidents are presented in Table 5. These results reveal that the MPA values achieved by all teams are remarkably higher than the *Best Performance Accuracy (BPA)* achieved by the top-performing teams in both subtasks. This suggests that the systems and approaches employed by all participants exhibit a high degree of complementarity. Additionally, when considering the diversity of errors among the participating teams, greater diversity is observed in Subtask 1 and in the detection of the "other" class in Subtask 2, this finding aligns with the improvement observed in the MPA metric of both subtasks.

	Class	BPA	MPA	CFD	
1	both	0.9375	0.9817	0.1615	8
2	accident	0.9757	0.9947	0.0517	7
2	murder	0.9713	0.9904	0.0668	7
2	robbery	0.9783	0.9939	0.0503	7
2	other	0.9323	0.9817	0.1559	7

Table 5: Comparison of BPA, MPA, and CFD results between the different general approaches. The first column refers to the subtask number while the last one refers to the number of systems' results involved in the calculation.

Qualitative Analysis

In order to further analyze the outcomes of the participating systems, we decided to take advantage of the obtained results in terms of MPA with the objective of identifying those instances that were misclassified by all the proposed approaches. A manual qualitative analysis was performed over a subset of these tweets. In the following paragraphs we briefly describe the main observed features as well as some samples of these instances. It is important to highlight that, usernames and proper names in the tweets have been replaced by a corresponding label in order to preserve the users' privacy. Similarly, the URLs have been also replaced.

Most of the tweets analyzed describe some actions performed by the authorities after a crime was committed, for example, a warrant for recovering stolen objects, the arrest of a murderer, or the discovery of corpses after an accident happened. These instances are easy to be misclassified since often they tend to contain similar vocabulary to the one in tweets reporting an incident. Table 6 shows some examples of these tweets. In this case, the three instances belong to the "negative" class for Subtask 1 and all systems identified as reporting a violent incident.

Tweet	Image
Mando Coordinado Policía Morelos detuvo a un hombre por homicidio •En el municipio de Temixco URL URL *Morelos' police coordinated command arrest a man by homicide • In the Temixco municipality URL URL*	Coordinación Estatal para la Construcción de la Paz Carlos Eduardo "N"
Detienen a elemento de la Policía vial de #Cuernavaca por posesión de un vehículo con reporte de robo #Morelos URL *Highway police officer from #Cuernavaca is arrested by holding a car with theft report #Morelos URL*	
—#ActualidadDL— Identifican muertos en accidente de la carretera Turística URL #DiarioLibre #Accidente #Carretera #Suceso #Santiago URL *—#ActualidadDL— Dead people on the accident in the touristic highway were identified URL #DiarioLibre #Accidente #Carretera #Suceso #Santiago URL*	

Table 6: Samples of tweets incorrectly classified due to describe some actions performed by the authorities.

We also identified some instances where the keywords related to a given violent incident are non-common words, for example in the first row of Table 7 which describes an homicide by means of only one keyword "ejecutado" (*murdered*) which is probably not wide used in the dataset. Indeed, this sample belongs to the "positive" class for subtask 1 and to the "murder" class for Subtask 2 while any system was able to correctly recognize these labels. The sample in the second row belongs to the "negative" class and all systems identified as reporting a violent incident. In this case, the tweet is about a murder committed during an alleged attempt of robbery, then it can be considered as reporting either of the violent incidents in the Subtask 2.

We also identified some instances that even labeled incorrectly by all systems maybe not considered as mistakes at all. For example, according to annotation guidelines, annotators were requested to not consider instances reporting accidents when they were about celebrities like singers, sports professionals, etc., since some of them can be generated with the intention of informing people more than reporting an incident. However, an instance like the one in the first row of

Tweet	Image
Con kiosco o sin kiosco, el tema es ese, la violencia en El Crucero. Hace unos minutos, un ejecutado en el parque 'La Corregidora', donde se llevan a cabo las obras de la segunda etapa del Mejoramiento de Imagen Urbana por parte del Gobierno del Estado: URL *With kiosk or without kiosk, this is the theme, the violence on El Crucero. Some minutes ago, a murdered in 'La Corregidora' park, where the works of the second stage of the Urban Image Improvement by the State government are carried out: URL*	
Asesinaron a un chofer de Uber en un presunto intento de robo en Ciudad Evita URL a través de @USER URL *An Uber driver was murdered in an alleged robbery attempt in Ciudad Evita URL via @USER URL*	

Table 7: Samples of instances incorrectly classified due to non-common words and ambiguities.

Table 8 was classified as positive due to it contains information about an accident while it belongs to the negative class. On the other hand, there are some samples like the second and third rows in the Table 8 which were labeled as positive for Subtask 1 (and even the third as "accident" for Subtask 2) when despite containing information about accidents, they are not reporting the event. This is likely to be an annotation mislead issue.

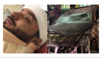
Tweet	Image
NAME, vocalista del grupo Valedores de la Sierra, sufrió un FUERTE accidente, se reporta como grave URL URL NAME, singer of the band Valedores de la Sierra suffered a STRONG accident, he is reported as severe URL URL	
Después del trágico accidente en la Sagarnaga en la ciudad de La Paz, la alcaldía se pronuncio al respecto y argumento que exigirán la devolución total de todos los gastos. #Terravisión URL *After the tragic accident in the Sagarnaga in the city of La Paz, the city hall spoke about it and argument that they will demand the full refund of all expenses. #Terravisión URL*	
@USER Este accidente cobró la vida de personas, sin embargo, los medios también deberían informar sobre todos aquellos hermanos que perecieron a consecuencia del #alcohol adulterado. Me duele ver a mi gente en peligro URL *@USER This accident took the lives of people, however, media should also communicate about over all those brethren who perished as a result of the adulterated #alcohol. It hurts to see my people in danger URL*	

Table 8: Samples of instances incorrectly classified due to annotation criteria.

4 Conclusions

We presented an overview of the DA-VINCIS 2023 shared task organized in the framework of IberLEF. DA-VINCIS 2023 promotes research into the identification of aggressive and violent incidents on social networks, a task of substantial societal significance. It used an upgraded version of the dataset from the previous edition of this task, encompassing the inclusion of at least one associated image per tweet, along with the reconsideration of categories and their manual annotation for the purpose of identifying violent incidents and their subcategorization. This evaluation campaign facilitated the assessment of a wide array of approaches, enabling a comparative analysis of their effectiveness. Various models, features, and techniques were presented within the proposed approaches, thereby contributing to the advancement of the field of identifying aggressive and violent incidents in the Spanish language.

The results indicate, as anticipated, that the fine-grained Subtask 2 presented greater difficulty compared to the binary classification one involved in Subtask 1. Notably, Transformer-based approaches exhibited a dominant presence and outperformed traditional machine learning methods employing BoW representations. The evaluation encompassed a stimulating array of proposals, introducing notable innovations such as data augmentation techniques facilitated by prompting engineering using the GPT-3 model. Additionally, back-translation was employed for enhancing the text modality, while approaches based on web searches were explored to obtain new relevant images, alongside image modification techniques. Furthermore, certain systems incorporated image captioning into the task, augmenting the contextual information available to the model for the identification of violent incidents. These aforementioned approaches collectively enhanced performance and effectively addressed the specific challenges inherent in the task at hand.

The inclusion of at least one associated image per tweet facilitates the utilization of both modalities for the purpose of identifying violent incidents on social media. The results derived from the proposed approaches by the participants underscore the benefits associated with leveraging both modalities, thereby highlighting their complementary nature. Building upon the success of this shared task, future work is proposed to extend the scope of aggression and violent incident detection to encompass videos disseminated on social networks. An additional challenge to be considered within this task involves the prediction and anticipation of locations where violent events are likely to occur.

Acknowledgements

This work was supported by CONACyT under grant CB-S-26314, *Integración de Lenguaje y Visión mediante Representaciones Multimodales Aprendidas para Clasificación y Recuperación de Imágenes* and scholarship 925996. We are also thankful with Sergio Nava for the support on the statistical analysis section.

References

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

Cabada, R. Z., M. L. B. Estrada, V. M. B. Beltrán, R. A. C. Sapien, N. L. López, G. Ángel Beltrán Ruiz, B. A. C. Sainz, and H. M. C. López. 2023. DA-VINCIS at IberLEF 2023: Detecting Aggressive and Violent Incidents from Social Media in Spanish using Text Information. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), CEUR Workshop Proceedings. CEUR-WS.org.*

Graff, M., S. Miranda-Jimenez, E. S. Tellez, and D. Moctezuma. 2020. Evomsa: A multilingual evolutionary approach for sentiment analysis [application notes]. *Comp. Intell. Mag.*, 15(1):76–88, feb.

Graff, M., D. Moctezuma, E. Tellez, and S. Miranda. 2023. INGEOTEC at DAVINCIS: Bag-of-Words Classifiers. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), CEUR Workshop Proceedings. CEUR-WS.org.*

Gutiérrez-Megías, A. J., F. Martínez-Santiago, L. A. Ureña-López, and A. Montejo-Ráez. 2023. SINAI participation at DA-VINCIS task in IberLEF 2023: Data augmentation for multimodal classification. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), CEUR Workshop Proceedings. CEUR-WS.org.*

Hernández-Minutti, B., J.-A. Olivares-Padilla, R. Valerio-Carrera, and O. J. Gambino. 2023. Detection of violent events in social media: DA-VINCIS 2023. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), CEUR Workshop Proceedings. CEUR-WS.org.*

Lin, T., Y. Wang, X. Liu, and X. Qiu. 2022. A survey of transformers. *AI Open*, 3:111–132.

Mata Rivera, M., M. Torres-Ruiz, G. Guzmán, R. Quintero, R. Zagal-Flores, M. Moreno, and E. Loza. 2016. 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*, 2016:1–11, 03.

Nava-Muñoz, S., M. Graff-Guerrero, and H. J. Escalante. 2023. Comparison of classifiers in challenge scheme. In *Pattern Recognition - 15th Mexican Conference, MCPR 2023, Tepic, Mexico, June 21-24, 2023, Proceedings*, volume 13902 of *Lecture Notes in Computer Science*, pages 89–98. Springer.

Pavao, A., I. Guyon, A.-C. Letournel, X. Baró, H. Escalante, S. Escalera, T. Thomas, and Z. Xu. 2022. CodaLab Competitions: An open source platform to organize scientific challenges. Technical report, Université Paris-Saclay, FRA., April.

Piña-García, C. and L. Ramírez-Ramírez. 2019. Exploring crime patterns in Mexico City. *Journal of Big Data*, 6, 07.

Ponce-León, E. and I. H. López-Nava. 2023. CICESE at DA-VINCIS 2023: Violent Events Detection in Twitter using Data Augmentation Techniques. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), CEUR Workshop Proceedings. CEUR-WS.org.*

Rubio, J. L. S., A. V. Almeida, and I. Segura-Bedmar. 2023. UC3M at Da-Vincis-2023: using BETO for Detection of Aggressive and Violent Incidents on Social Networks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), CEUR Workshop Proceedings. CEUR-WS.org.*

Sandagiri, C., B. Kumara, and B. Kuhaneswaran. 2020. Detecting Crime Related Twitter Posts using Artificial Neural Networks based Approach. pages 5–10, 11.

Tang, E. K., P. N. Suganthan, and X. Yao. 2006. An analysis of diversity measures. *Mach. Learn.*, 65(1):247–271.

Vallejo-Aldana, D., A. P. López-Monroy, and E. Villatoro-Tello. 2023. Enhancing Multi-modal Classification of Violent Events using Image Captioning. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), CEUR Workshop Proceedings. CEUR-WS.org.*