Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 20□□-2 recibido 05-07-2022 revisado 22-07-2022 aceptado 25-07-2022

Overview of DA-VINCIS at IberLEF 2022: Detection of Aggressive and Violent Incidents from Social Media in Spanish

Resumen de la Tarea DA-VINCIS en IberLEF 2022: Detección de Incidentes Violentos en Redes Sociales en Español

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.
{arellano.luis, hugojair, villasen, mmontesg}@inaoep.mx
fernando.sanchez@cimat.mx

Abstract: This paper presents the overview of the DA-VINCIS 2022 task, organized at IberLEF 2023 and co-located with the 38th International Conference of the Spanish Society for Natural Language Processing (SEPLN 2022). DA-VINCIS challenged participants to develop automated solutions for the detection of violent events mentioned in social networks. We released a novel corpus collected from Twitter and manually labeled with 4 categories of violent incidents (plus the no-incident label). The shared task focused on the Mexican variant of Spanish and it was divided into two tracks: (1) a binary classification task in which users had to determine whether tweets were associated to a violent incident or not; and (2) a multi-label classification task in which the category of the violent incident should be spotted. More than 40 teams registered for the task and 12 participants submitted predictions for the final phase. Very competitive results were reported in both sub tasks, where transformer-based solutions obtained the best results. Corpora and results are available at the shared task website at https://codalab.lisn.upsaclay.fr/competitions/2638.
Keywords: DA-VINCIS, violent event detection, text classification.

Resumen: Se presenta el resumen de la tarea DA-VINCIS 2022, organizada en IberLEF 2022 junto a la 38ª Conferencia Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2022). DA-VINCIS plantea el reto de detectar automáticamente piezas de información en redes sociales que estén asociadas a eventos violentos. Se liberó un nuevo corpus para el Español Mexicano que fue etiquetado manualmente con 4 categorías de eventos violentos (además de la categoría no-violento). Se propusieron dos subtareas: (1) una tarea de clasificación binaria donde se buscaba distinguir tuits asociados a eventos violentos del resto; y otra (2) donde se buscaba identificar la categoria del evento violento. Más de 40 participantes se registraron en el portal y 12 enviaron resultados para la fase final. Los resultados obtenidos fueron muy competitivos para ambas tareas; las soluciones que obtuvieron los mejores resultados se basaron en modelos tipo *transformer* para el español. El corpus y los resultados detallados pueden consultarse en el sitio web de la tarea: https://codalab.lisn.upsaclay.fr/competitions/2638.
Palabras clave: DA-VINCIS, Detección de eventos violentos, Clasificación de textos.

ISSN 1135-5948. DOI 10.26342/2022-69-18

1 Introduction

Violence has obvious negative effects on those who witness or experience it, including a higher incidence of depression, anxiety, post-traumatic stress disorder, among others. In addition, violence events have a high impact for governments, as they are in charge of guaranteeing security to their population. Therefore, the detection and tracking of violence related events is critical. In this context, social networks comprise a valuable information source for the detection and monitoring of violent events, as people very often post publications notifying the occurrence of violent events in real time. This represents an important opportunity for IT researchers that can provide solutions based on natural language processing for the timely detection of violent incidents in social networks. Solutions of this kind could be used by authorities to respond more efficiently to events happening in real time, and to develop crime prevention policies according to geographical zones and types of events. Likewise, such solutions would be very helpful to the population, as one could know what violent events are happening in which zones in real time.

We organized a shared task collocated with IberLEF2022 called DA-VINCIS. This task focused on the detection of violent incidents on Twitter. It challenged participants to develop methods able to classify tweets as reporting a violent event or not. For this first edition, the shared task targeted Spanish in its Mexican variant. This is motivated by the lack of resources in Spanish for approaching the task, and the fact that Mexican Spanish is the most spoken variant of this language[1]. We released a novel corpus carefully labeled according to violent event categories. The shared task comprised two tracks: *violent event identification* and *violent event category recognition*. Labeled data was provided to participants for both tracks for the development of their solutions, and unlabeled data was used for the final evaluation of the corresponding tracks.

As far as we know, this was the first shared-task aiming at detecting violent events from social media. This is an issue that has received little attention from the community, despite its enormous potential impact. Therefore, the aim of the challenge was to motivate research on a topic little explored in Spanish, but with great potential impact for the whole population and authorities. In addition, an implicit goal was to raise awareness of the relevance of this problem.

The task posed several challenges to the community, including: dealing with Mexican Spanish, the ambiguous language inherent to Twitter, the high class imbalance ratios present in our datasets, among others. We are confident that the shared task will give rise to novel solutions that could be used in the near future for applications of societal impact, for example, generating real-time occurrence crime maps. Last but not least, we plan to release the associated corpus in the near future so that the community can keep working on it even at the end of the shared task.

The remainder of this paper is organized as follows. Section 2 describes the shared task in detail. Then, Section 3 introduces the DA-VINCIS corpus. Section 4 presents the results obtained and a summary of participants' solutions. Finally, Section 5 outlines conclusions and future work directions.

2 Task description

As previously mentioned, the DA-VINCIS shared task comprised two tracks: a binary classification subtask that aimed at distinguishing tweets associated to violent incidents from those that are not; and (2) a task that challenged participants to identify the type of violent incident (if any) being reported in tweets. The categories considered for the latter task are described in Table 2.

The DA-VINCIS corpus, described in detail in the next section, was used for the evaluation of both subtasks. The challenge was run in the CodaLab platform (Pavao et al., 2022). The shared task was divided into two stages as follows:

- **Development phase.** Participants were provided with labeled training data and unlabeled validation data. During this phase, which lasted about two months, participants were able to submit predictions for the validation set and receive immediate feedback in the CodaLab site.
- **Final phase.** Participants were provided with unlabeled test data. They were able to upload up to five submis-

[1] In terms of the number of native speakers.

Reference	Considered categories
(Mata Rivera et al., 2016)	Theft, Crime, Theft with violence, Theft walking, Theftin car, Theft without violence
(Sandagiri, Kumara, and Kuhaneswaran, 2020b)	Assault, Burglary, Drugs Violations, Homicide, Sex Offences, Suicide
(Sandagiri, Kumara, and Kuhaneswaran, 2020a)	Assault, Burglary, Drugs Violations, Homicide, Sex Offences
(Piña-García and Ramírez-Ramírez, 2019)	Robbery passerby, Theft of motor vehicle, Robbery of business property, Card fraud, Homicide, Domestic burglary, Robbery on public transportation, Rape, Firearm injuries, Robbery in subway, Robbery on taxi, Robbery to Carrier, Robbery to deliver person
DA-VINCIS	Accident, Homicide, Theft, Kidnapping, Non-incident

Table 1: Violent incidents considered in previous work.

sions during the competition. Performance on the test set was used to rank participants.

For subtask 1, recall, precision and f_1 score with respect to the *violent-incident* class were considered as evaluation measures. For subtask 2, macro average recall, precision and f_1 score were considred. In both cases, the leading evaluation measure was that of f_1 score.

3 DA-VINCIS corpus

The DA-VINCIS corpus is a collection of tweets[2] associated to reports of violent incidents in Mexican Spanish. The aim of this novel corpus is to boost research in the automated detection and monitoring of violent incidents in social networks. Summarizing, a large number of tweets was retrieved using queries associated to predefined categories. Then, the tweets were filtered, and a subset of these was manually labeled. In the remainder of this section we provide details on the construction of this corpus.

A set of categories of violent incidents was defined after a careful analysis of relevant literature, see Table 1. The categories considered in each study differ according to the legal, psycho-social or geographical context, and commonly they are finally filtered by the criteria of the research group involved.

[2]Please note that the corpus is formed by both, the text in tweets and their associated images, if any, for this shared task only textual information was considered.

The categories considered in the DA-VINCIS corpus are shown in the last row of Table 1. The criteria for selecting such categories involved: categories that appear in most of previous studies (e.g., *Homicide* and *Theft*), generic categories (e.g., we considered a single Theft category) and categories that appeared most frequently among in Twitter accounts associated to local news in Mexico (e.g., *Kidnapping*). Finally, our choice for these categories relied on the number of tweets that we retrieved per each category.

It is important to mention that since the long term goal of this project is the real-time monitoring of violent incidents, the *Accident* category was taken into account. As on a daily basis authorities use the same communication channels to deal with this kind of problem. Categories such as *Sexual offences* and *Drug violations* were initially considered because of their relevance and the urgent need to prevent them. However the study of these categories is particularly complicated, because although there are reports or complaints on the internet, these are not frequent, in addition to the fact that these are common topics of conversation and discussion, therefore it makes it extremely difficult to find the reports among the large number of opinions. Definitions for the categories considered in the DA-VINCIS corpus are shown in Table 2.

To obtain keywords for the retrieval of violent incidents tweets, a research work was carried out where 30,000 tweets published in news accounts in Spanish were recovered, 5,000 tweets were manually tagged to identify if the news was violent (i.e. binary labeling) once established, an ML model was applied to label the rest of the corpus, the pseudo labels obtained were used to study the tweets and the unigrams, bigrams and trigrams that provided the most information for the classification, the most significant words from the top-100 were filtered, these were the keywords used to search for tweets of violent incidents.

Once the keywords were obtained, a tweet retrieval was performed using each of the selected keywords, where it was required that (1) tweets had an associated image, (2) language was Spanish and, (2) the tweet was geolocated in approximation to Latin America. The result of this process were 8000 tweets that were further filtered by eliminating those

Category	Definition
Accident	Eventual event or action that results in involuntary damage to people or things.
Homicide	Deprivation of life.
Theft	Seizure or willful destruction of someone else's property without the right and without the consent of the person who can legally dispose of them.
Kidnapping	Deprivation of liberty.
Non-incident	Selected when there is no crime reported.

Table 2: Definition of the categories considered in the DA-VINCIS corpus.

that could no longer recover any of their elements, that were written in a language other than Spanish but that were filtered in the search and the empty elements or that only consisted of a series of hashtags.

The filtered dataset was formed by 5000 tweets. Each tweet in the dataset was labeled by at least two annotators. Labels assigned by annotators took into account the context provided by the text of the tweet and the associated images, if any. However, even when having all the context available, the labeling process was not straightforward in some cases. Sometimes the images were conducive to confusion or vice versa. For example, confusing a traffic accident with a homicide with a car (cyclist hit by a car).

Randomly selected tweets from each category are shown in Table 3. Despite this samples were assigned the correct label, there were some samples that could be considered noisy, see Section 4. Table 4 shows the proportion of samples per class in the dataset. Please note that since categories are not necessarily disjoint (except the no-incident one). More than one label can be assigned to a single tweet, that is why the total number of labels is different from 5000.

To analyze the difficulty of the task, the agreement between the annotators was calculated. Table 5 presents the results of the Kappa coefficient, by the number of judgments collected. The coefficient values indicate moderate agreement, see (McHugh, 2012). However, we found some samples with noisy annotations, see Section 4, evidencing the need for a detailed curation for the DA-VINCIS corpus.

4 Participants approaches and results

In the following subsections we describe the main ideas addressed by the different participants, and present a general analysis of their results.

4.1 Systems' descriptions

A total of 12 teams participated in the DA-VINCIS shared-task; the majority tackled both subtasks, however, four teams only addressed the subtask 1 and one team only presented a solution for subtask 2. Something interesting to highlight is that the teams with the best performance in each of the two subtasks presented a proposal exclusively focused on the particular subtask. This indicates that both subtasks have their own specific challenges and, therefore, that it is not always convenient to approach them using the same strategy.

From the different solutions presented at the DA-VINCIS shared task, we found several coincidences, which indeed align with some general trends in Natural Language Processing. The main shared aspects correspond to the use of:

- **Pretrained Transformers**: All participant approaches used some pretrained transformer to take advantage of all the knowledge encoded in their pretraining. Some applied the traditional fine-tuning (Ta et al., 2022b), while others proposed some interesting modifications (Vallejo-Aldana, López-Monroy, and Villatoro-Tello, 2022; Turón et al., 2022; Montañés-Salas, del Hoyo-Alonso, and Peña-Larena, 2022; García-Díaz et al., 2022; Ta et al., 2022b). On the other hand, some approaches used the pretrained transformer as a frozen source of knowledge, only using the contextual embedding encodings (García-Díaz et al., 2022), or extracting relationships from the instances and task description using a Prompt-based framework (Qin et al., 2022).
- **Ensembles**: Multiple approaches employed ensembles to take advantage of variations of their base solution models. For example, the majority voting scheme was successfully used for subtask 1 (Vallejo-Aldana, López-Monroy, and Villatoro-Tello, 2022; Turón et al., 2022;

Categories	Original text	Translation
Accident	*#Ahora Reportan accidente de tránsito en el ingreso al municipio de Salcajá. Dos vehículos tipo picop involucrados en el percances. Precaución al conducir por el sector. Ampliaremos la información. #Stereo100Noticias*	#Now Car accident is being reported at the entrance of the Salcajá municipality. Two pickup vehicles involved. Caution when driving nearby. We will extend the information #Stereo100Noticias
Homicide	*La violencia y las ejecuciones continúan cada día en la CDMX un hombre fue ejecutado a 2 calles de la alcaldía de Cuahutémoc en la calle de Pedro Moreno*	Violence and killings continue everyday in CDMX a mean was killed two blocks from Cuahutémoc town hall in Pedro Moreno street
Theft	Imágenes en las que un sujeto que ingresó a robar a un local ubicado en Av.Tonalá y Madero en la Cabecera Municipal. El hombre iba armado y después del robo huyó en un auto Kia color gris que lo esperaba afuera del local.	footage in which a subject that entered to steal a facility in Av. Tonalá and Madero in the municipality. The man was armed and after the robbery escaped in a grey Kia that was waiting outside the facility.
Kidnapping	Secuestraron a sujeto frente al palacio municipal de Coatzacoalcos A plena luz del día realizan acto delictivo; los detienen y desarticula UECS banda de plagiarios recién formada; se quedan en el Cereso Duport Ostión	A man was kidnapped in front of Coatzacoalcos' town all. The criminal act was perfored during daylight; they were arrested and the UECS dismantled a band of kidnappers just formed; they are staying in the Duport Ostión prision.

Table 3: Samples from the DA-VINCIS corpus for the violent incident categories.

Montañés-Salas, del Hoyo-Alonso, and Peña-Larena, 2022). More sophisticated ensemble techniques were also applied, such as intermediate fusion of NNs using Knowledge Integration (KI), ensemble learning (García-Díaz et al., 2022), and a kind of multilevel fusion that incorporates information from multiple sources (Qin et al., 2022).

- **Multi-Task Learning (MTL)**: Several proposals took advantage of the pairing of subtasks 1 and 2 to carry out some kind of multi-task learning. For example, (Vallejo-Aldana, López-Monroy, and Villatoro-Tello, 2022) performed MTL for subtask 1 through a binary transformation of subtask 2 that is jointly learned in order to incorporate additional information for subtask 1. In contrast, in (Ta et al., 2022b) the prediction of each class of subtask 2 was transformed into a binary problem that is performed with MTL on the complete set of binary problems. Finally, (Ta et al., 2022a) proposed an interesting MTL approach where subtask 2 was carried out while jointly learning to distinguish real instances from instances generated by a GAN.

Categories	# Examples	% Total
Accident	1800	33.45
Homicide	417	7.75
Theft	286	5.31
Kidnapping	72	1.33
Non-violent	2878	53.49

Table 4: Proportion of samples from each class

Judgments	Tweets	Coefficient
2	1349	0.5758
3	1643	0.5767
4	1024	0.5979
5	350	0.5829

Table 5: Kappa coefficients by number of judgments (only results for 2 to 5 judgments are shown).

- **Data Augmentation (DA)**: This technique was also widely used by the participant teams. The most used method was back-translation (Turón et al., 2022; Montañés-Salas, del Hoyo-Alonso, and Peña-Larena, 2022; Ta et al., 2022b; Ta et al., 2022a), however, some approaches also integrated the examples in the intermediate languages to the augmented data, and in consequence used multilingual models in their training phase (Tonja et al., 2022).

- **Preprocessing**: Most teams performed standard preprocessing operations to allow the transformers-based language

models to handle the input texts. For example, they removed URLs, hashtag symbols, non-alphanumeric symbols, and adjusted user mentions (strings with @). Additionally, in (Vallejo-Aldana, López-Monroy, and Villatoro-Tello, 2022; Turón et al., 2022; Montañés-Salas, del Hoyo-Alonso, and Peña-Larena, 2022) emojis were replaced by their descriptive words, and acronyms and abbreviations were expanded in (García-Díaz et al., 2022).

Three approaches show some interesting features that do not fit the generalities described above; these are:

- **Noise Reduction**: *VICOMTECH* (Turón et al., 2022) carried out a relabelling process of the training data considering the votes of 5 systems learned from the original noisy data set. They "corrected" the instance labels if at least 4 of the 5 systems agreed to do so; using this approach they modified around 5% of the training set labels.
- **Use of Advanced Linguistic Features**: *UM-UJ-URJC* (García-Díaz et al., 2022) considered the use of a variety of features with the purpose of taking into account multiple aspects of the writing and communication style of tweets.
- **Use of Prompt Learning:** GDUT (Qin et al., 2022) employed a prompt learning module to inject information from a pre-trained language model into the violent event category recognition task. This approach incorporates the text provided by the prompt module into the tweet representation.

4.2 Evaluation campaign results

Table 6 presents the results obtained by the participant teams in subtask 1, the binary identification of violent incidentes. The teams are sorted by their F1-score over the positive class (i.e., the violent incident class); Precision and Recall are also reported to allow a better interpretation of these results. At the bottom it is included our baseline[3]

[3]Please note that during the final phase of the shared task we uploaded a single run of the baseline that obtained better results. However, in this paper we report the average over 10 runs of the performance of the baseline, which is a more reliable estimate of its performance.

Subtask 1: Binary violent event identification			
Team	**Precision**	**Recall**	**F1-Score**
CIMAT-UG-UAM-IDIAP	0.803	0.750	**0.775**
VICOMTECH	**0.812**	0.737	0.773
ITAINNOVA	0.779	0.751	0.765
UM-UJ-URJC	0.774	0.753	0.764
Sdamian	0.761	0.750	0.756
Bernardo	0.780	0.730	0.754
IPN-DLU-UNOMAHA-1	0.755	0.740	0.748
CIC-IPN	0.761	0.730	0.745
IPN-DLU-UNOMAHA-2	0.740	0.747	0.744
JuanCalderon	0.723	**0.763**	0.742
Sustaitangel	0.710	0.742	0.726
Baseline	0.763	0.780	0.750

Table 6: Results of the participant teams in Subtasks 1. They correspond to the Precision, Recall and F1 score in the positive class.

result, which corresponds to the direct use of a traditional fine-tuning (i.e., using a single linear layer and the use of softmax for classification) of BETO (Cañete et al., 2020), a well-known pre-trained language model in Spanish.

The best performance in Subtask 1 was obtained by the CIMAT team (Vallejo-Aldana, López-Monroy, and Villatoro-Tello, 2022) followed by VICOMTECH (Turón et al., 2022) These two approaches have in common that they took advantage of the parallelism of both subtasks, particularly, they included in their model for subtask 1 some information from subtask 2. The third best approach is ITAINNOVA (Montañés-Salas, del Hoyo-Alonso, and Peña-Larena, 2022), which, similarly to the CIMAT team, used an ensemble of multiple transformer-based models. On the one hand, the CIMAT approach combined the output of three BERT-based models fine-tuned to perform MTL. In this case, MTL is used to simultaneously learn the subtask 1 and a binary version of a violent event subcategory classification (i.e., each BERT model is different in the specific event subcategory chosen). On the other hand, the ITAINNOVA approach uses different pretrained models (such as BETO, Twitter-XLM-Roberta and BSC-Roberta), thus obtaining its diversity from the models and not from the data.

It should be noted that the different approaches obtained very close results; the best performance is only 6.8% greater than the lowest, and the standard deviation of the set of F1-scores is only 0.015.

The results obtained by the teams in subtask 2, the violent event category recognition, are shown in Table 7. The best performance in this subtask corresponds to the GDUT

Subtask 2: Violent event category recognition			
Team	**Precision**	**Recall**	**F1-Score**
GDUT	0.550	**0.564**	**0.554**
VICOMTECH	0.517	0.545	0.528
ITAINNOVA	0.509	0.503	0.504
CIC-IPN	0.467	0.520	0.490
CIMAT-UG-UAM-IDIAP	**0.655**	0.421	0.473
UM-UJ-URJC	0.442	0.549	0.469
Sustaitangel	0.459	0.424	0.433
CIC-IPN-DLU-UNOMAHA-1	0.377	0.438	0.392
Baseline	0.498	0.460	0.570

Table 7: Results of the participant teams in the Subtask 2. They correspond to the macro average values of Precision, Recall and F1 score.

team. Their solution is mainly characterized by the incorporation of semantic relations between the text instances and the name of the categories through the application of a Prompt learning module. The second and third best performances were obtained, as in subtask 1, by the VICOMTECH (Turón et al., 2022) and ITAINNOVA (Montañés-Salas, del Hoyo-Alonso, and Peña-Larena, 2022) teams, respectively. Their adequate performance in both tracks suggests the relevance and robustness of these two approaches for the task addressed.

From the results, it is notorious that subtask 2 is much more challenging than subtask 1; something expected due to the high imbalance in some of the categories. This is reflected in a greater standard deviation (0.051) in the reported F1-scores, and also in the larger difference between the best and worst reported results (in this case, the former is 41% greater than the later).

4.3 Analysis

To provide insights on the complimentary and redundancy of solutions, the Intra-ensemble Coincident–Failure Diversity (CFD) was calculated for the 11 submissions ranked in Table 6. This index indicates how diverse the errors of each model are with respect to each other if one would build an ensemble with them. The resulting CFD was 0.5590, indicating regular diversity (the range of values of CFD is [0,1]), that could be exploited for building a more robust model. On the other hand the maximum possible accuracy (a tweet is counted as well classified, if any of the models classified it correctly) was 0.9181. Further evidencing the potential benefits of building an ensemble with the 11 evaluate solutions.

In order to illustrate the inherent difficulties of the shared task, Table 8 shows examples of tweets that were missclassified by most participants when approaching subtask 1. Several interesting aspects can be discussed around these examples. First, there are tweets that were wrongly labeled by the annotators, for instance sample 3. This was a problem highlighted by participants during the challenge (Turón et al., 2022). Secondly, there are tweets for which the assigned category is debatable. For instance, tweet 4 refers to an accident happening in the context of an F1 race, it is an accident, but not really relevant for the purpose of the project. Also, tweet 1 refers to a report associated to several violent events happening in different places (we hypothesize this is why it was labeled as Non-violent). Summarizing, a large portion of samples missclassified by the systems could be due to subjective labeling. Therefore, we conclude the dataset needs of further manual curation. Still, we think the DA-VINCIS corpus is a valuable resource that will boost research in this relevant task.

5 Conclusions

The DA-VINCIS shared task at IberLEF promotes research into the identification of violent incidents on social networks, a task with a high social impact. A new dataset for the task of identification of violent incidents as well as their subcategorization is presented. This evaluation campaign made it possible to evaluate an important diversity of approaches and contrast their effectiveness. Different models, characteristics and techniques of the proposed approaches were presented, contributing to the progress of the identification of violent incidents in Spanish language.

The results indicate, as might be expected, that the fine-grained subtask 2 was more challenging. A strong presence of approaches based on transformers was found, but also there was a vitalizing variety of proposals with important novelties such as the application of GANs, the automatic correction of instances, and the use of non-learning tools to act as a kind of oracle, all of them introduced to improve the methods' performance as well as to to deal the specific challenges of the task at hand.

It was found that having some information on the subcategory of the general class of interest seems to help to make a better iden-

ID	Translation	Text	Category
1	Intense police activity in Coacoatzintla's municipality, in response to the supposed kidnap of a young male. A family member of the kidnapped person was killed when trying to impede this crime. In Jilotepec was found the vehicle where the person was abducted SP_Veracruz	*Una fuerte movilización policiaca se registró en el municipio de Coacoatzintla ante el presunto secuestro de un joven. Al tratar de impedir el hecho, un familiar fue asesinado. En Jilotepec fue hallado el vehículo en el que se cometió el ilícito SP_Veracruz*	Non-violent
2	30 years now from the Cimitarra massacre, a violence act that left more than 250 thousand deaths turning Colombia into a a huge common grave.	*A 30 años de la masacre de Cimitarra, una violencia que dejo más de 250 mil muertos convirtiendo a Colombia en una gran fosa común.*	Violent
3	Homicide - In a clinic at #Cartago Bibiana Liseth Guzmán Ordóñez, 31 years old and official of the @ipscartago, died, after she was shoot with a firearm. In the same incident a 26 years old man was hurt. The women left a daughter.	*Homicidio - En una clínica de #Cartago falleció Bibiana Liseth Guzmán Ordóñez de 31 años de edad, funcionaria de la @ipscartago luego de que le propinaran varios impactos con arma de fuego. En este mismo hecho resultó lesionado un hombre de 26 años. La mujer deja una hija.*	Non-violent
4	"The accident could have been avoided if they would leave me enough space to take the curve. You need of two persons for this to work, and I felt they throw me away. When we challenge to each other in a race this things can happen, unfortunately."	"El accidente se pudo haber evitado si me hubieran dejado espacio suficiente para tomar la curva. Necesitas 2 personas para que esto funcione y yo sentí que sacaban. Cuando nos retamos mutuamente en una carrera estas cosas pueden pasar, desafortunadamente."	Violent

Table 8: Examples of tweets incorrectly classified by all of the participant teams.

tification. Multitask Learning is strongly positioned as a good alternative that improves performance and takes advantage of the parallelism between subtasks 1 and 2. These findings open the possibility that other future approaches could use virtual subtasks with different fine grain levels in order to take advantage of this type of scheme.

Likewise, an in depth analysis of the corpus revealed that there is room for improvement in terms of the quality of annotations. On the one hand, a curation process trying to identify noisy annotations should be performed. On the other hand, the definition of categories should be further tuned, so that annotation guidelines result in objective annotations. This is work in progress.

As previously mentioned, the DA-VINCIS corpus also comprises visual information, therefore another venue of current work is studying the potential added value of using images associated to tweets when detecting violent incidents. The corpus will allow us to study the performance of solutions that consider multimodal information.

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.* We also would like to thank CONACyT for partially supporting this work under grant CB-2015-01-257383. Additionally, the authors thank CONACYT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies.

References

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

García-Díaz, J. A., S. M. Jiménez-Zafra, M. Rodríguez-García, and R. Valencia-García. 2022. UMUTeam at DA-VINCIS 2022: Aggressive and Violent classification using Knowledge Integration and Ensemble Learning. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

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.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic.

Montañés-Salas, R. M., R. del Hoyo-Alonso, and P. Peña-Larena. 2022. ITAINNOVA@DA-VINCIS: A Tale of Transformers and Simple Optimization Techniques. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

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.

Qin, G., J. He, Q. Bai, N. Lin, J. Wang, K. Zhou, D. Zhou, and A. Yang. 2022. Prompt Based Framework for Violent Event Recognition in Spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

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

Sandagiri, S., B. Kumara, and B. Kuhaneswaran. 2020b. Deep Neural Network-Based Approach to Identify the Crime Related Twitter Posts. *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pages 1000–1004.

Ta, H. T., A. B. S. Rahman, L. Najjar, and A. Gelbukh. 2022a. GAN-BERT: Adversarial Learning for Detection of Aggressive and Violent Incidents from Social Media. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

Ta, H. T., A. B. S. Rahman, L. Najjar, and A. Gelbukh. 2022b. Multi-Task Learning for Detection of Aggressive and Violent Incidents from Social Media. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

Tonja, A. L., M. Arif, O. Kolesnikova, A. Gelbukh, and G. Sidorov. 2022. Detection of Aggressive and Violent Incidents from Social Media in Spanish using Pretrained Language Model. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

Turón, P., N. Perez, A. García-Pablos, E. Zotova, and M. Cuadros. 2022. Vicomtech at DA-VINCIS: Detection of Aggressive and Violent Incidents from Social Media in Spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

Vallejo-Aldana, D., A. P. López-Monroy, and E. Villatoro-Tello. 2022. Leveraging Events Sub-Categories for Violent-Events Detection in Social Media. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*