

Enhancing Multi-modal Classification of Violent Events using Image Captioning

Daniel Vallejo-Aldana^{1,*}, Adrián Pastor López-Monroy¹ and Esaú Villatoro-Tello²

¹*Department of Computer Science, Mathematics Research Center (CIMAT), Guanajuato, Mexico*

²*Idiap Research Institute, Switzerland*

Abstract

This research paper presents our involvement in the collaborative evaluation campaign of DA-VINCIS@IberLEF 2023. Our focus lies on tackling the Violent-Event Identification (VEI) task, wherein we employ a multi-modal approach that combines textual input with image captions extracted from visual data. The obtained results demonstrate a competitive performance, as we achieved the top position in the VEI task with an average F1 score of 0.92638.

Keywords

Multi-modal models, Natural Language Processing, Data Oversampling, Parameter Tuning, Image Captioning

1. Introduction

Social media has become in recent years an integral part of journalistic work [1]. It helps to intimate journalists with the audience by delivering fast and accessible information to everyone. In the work elaborated by [2] Twitter has become one of the main social media platforms to deliver information from citizen journalists due to the platform amenities. These amenities encompass the ability to tag, link, and address specific individuals, as well as to forward messages from others in order to disseminate information among diverse groups sharing common environments and interests [2]. Most of the information shared among Twitter users may not contain any relevant information related to security interests. However, citizen journalism may help to detect relevant security events faster than traditional journalism methods such as television and newspapers at the cost of having biased information due to the subjective factor of the writer.

The objective of the DA-VINCIS@IberLEF challenge [3] is to discern pertinent information for the identification (VEI - Violent-Event-Identification) and categorization (VEC - Violent-Event-Categorization) of violent events sourced from social media. During the 2022 edition of the DA-VINCIS@IberLEF [4] challenge, only textual information was provided as input data for the VEI and VEC sub-tasks. In order to address the VEI shared task [5] suggest a

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.


✉ daniel.vallejo@ciamat.mx (D. Vallejo-Aldana); pastor.lopez@ciamat.mx (A. P. López-Monroy);

esau.villatoro@idiap.ch (E. Villatoro-Tello)

🌐 <https://github.com/danielvallejo237> (D. Vallejo-Aldana)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

multitask approach that captures pertinent information, effectively combining the VEI and VEC tasks into a single model achieving the best results in Violent-Event-Identification (VEI). In the VEC task, the utilization of a prompt-based approach, as detailed by [6], yielded the highest performance. Additionally, in the Violent-Events Categorization shared task, strategies such as data augmentation through back-translation [4], noise-reduction [7] and model ensembles [5] were employed to enhance the achieved results.

In the 2023 edition of the DA-VINCIS@IberLEF [4] challenge, a set of images accompanied by corresponding social media text is presented as the input data. In this work, we want to explore how to use visual and textual information together to create a multi-modal approach to solve both sub-tasks. We will explore the importance of parameter tuning during training to obtain better results at the inference stage. In this research paper, we propose to use modern Transformer architectures like RoBERTa [8] to address both sub-tasks, extracting important features from the images. To accomplish this task, we employ BLIP [9], a pre-trained model specifically designed for image captioning. Subsequently, we merge these generated captions with their respective text counterparts using a designated text separator. We would like to see how visual information complements textual information and helps to improve classification performance. To boost the classification performance of the model we use data oversampling, a weighted loss function, and an ensemble configuration. Our proposal obtained first place in the Violent-Event-Identification (VEI) shared task (sub-task 1) with an F1 score of 0.92638 and an F1 score of 0.84207 for Violent-Event-Categorization (VEC) shared task (sub-task 2).

2. Task Description and Data

The DA-VINCIS@IberLEF 2023 [3] is composed of two tasks: (1) Violent-Event-Identification (VEI) which consists in detecting whether the input data (text and image) contains information about a violent event and (2) Violent-Event-Categorization (VEC) aiming to detect the violent event sub-type (*Traffic Accident, Murder, Robbery, Other*). The training dataset comprises 2996 examples; for sub-task 1 we have 1277 positive examples and 1719 negative examples. For sub-task 2 we have the categories percentages shown in Table 1.

Violent-Events Categories	Categories Percentage
Traffic Accident	31.38 %
Murder	6.01 %
Robbery	5.24 %
Other	57.38 %

Table 1

Categories percentage for sub-task 2, we see that categories *Murder* and *Robbery* are under-represented making the training data set a highly imbalanced data set.

From Table 1 we observe that sub-task 2 is quite challenging due to the low number of examples of classes *Murder* and *Robbery*. The Violent-Event-Categorization task is designed to be a multi-label task, however, the number of input texts that belong to multiple classes is around 1% of all the dataset. Hence we decided to treat the VEC problem as a multi-class classification task. Each text example is associated with one or multiple images related to the

tweet content, for the training data set we have 4259 images for all the text examples.

3. Methodology

To determine how to use the visual and textual information in a way that can be used to accurately identify and categorize violent events, we propose multiple multi-modal approaches either combining the outputs of the model's representation vectors or by joining textual descriptions of each one of the images related to a text. To extract the important information from the images, we propose different approaches such as using a Convolutional Neural Network (CNN) or a pre-trained image captioning model such as BLIP [9]. With a viable multi-modal setup, our proposition entails the utilization of data oversampling, a weighted-loss function, and a model ensemble configuration. These subtle modifications in model training have the potential to substantially enhance the model's performance.

3.1. Image Feature Extraction

- **Convolutional Neural Networks:** Our first approach to extracting important features from the images is to use a Convolutional Neural Network (CNN), for this purpose, we used a pre-trained version of Inception-v3 [10], on the IMAGENET [11] data set. We then fine-tune this model to either detect or categorize violent events. To create a single image per text, we concatenate the corresponding images for each tweet to form a final image that is fed to the Convolutional Neural Network. The image pre-processing steps applied to the image are the ones described in [10].
- **Image Captioning:** We evaluate another approach to extracting important information from the visual data using image captioning. To this aim, we used a BLIP[9] pre-trained model applied to each one of the images obtained from the training data. This generates a caption describing each one of the images. To use this information in the same language as the tweets, we use a pre-trained Marian Neural Machine Translation Model [12] to translate the captions from English to Spanish. To join the information obtained from different images we use the connecting word *and* (*y* in Spanish) to generate a single sentence for each one of the input texts. To correct some of the generated captions we use regular expressions to eliminate repeating words one after the previous one. An example of the connected sentence and the correction using regular expressions is shown below.



Original obtained caption: two ak ak ak ak ak ak ak ak ak ak ak ak ak ak ak ak and a man with a bald haircut and a bald face
Caption after correction with regular expressions: two ak and a man with a bald haircut and a bald face.

The histograms depicted in Figure 1 display the lengths of tweets and captions in the training dataset. It is evident that both tweets and captions are relatively short, thereby falling

well within the text-length limitations of the Transformer model [13] (approximately 512 tokens) and allowing for easy handling.

Lengths of tweets (on the left) and captions (on the right) in the training dataset

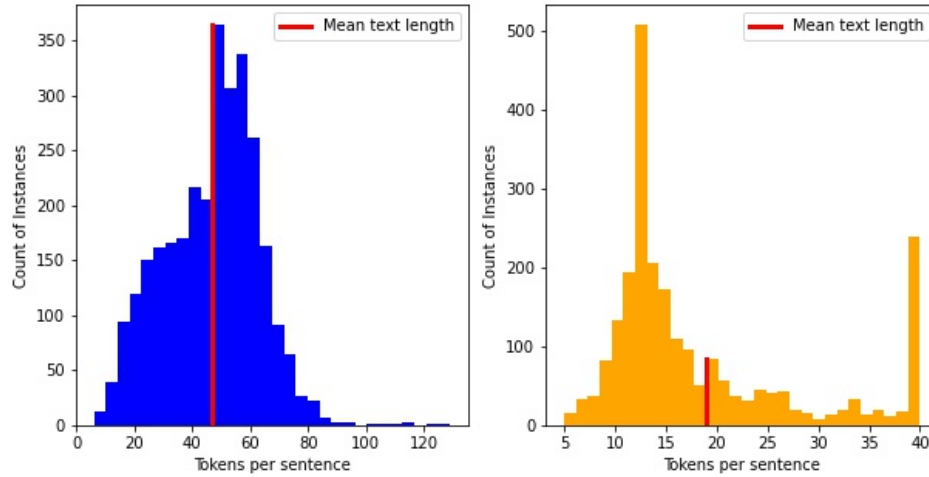


Figure 1: Lengths of tweets (blue histogram) and captions (orange histogram) in the training dataset. The average length of a tweet is around 47 tokens whereas the average length of an image caption is around 18 tokens.

3.2. Multi-Modal Approaches

The textual information contains the most relevant information about violent events identification and categorization (See EXP6 Section 4.1.2). We use a RoBERTa [8] model using the base configuration (Embedding dimension of 768) using pre-trained weights within a Spanish tweet domain described in [14]. To increase the information related to an event, we incorporate the visual information with different multi-modal approaches using the features from a CNN or from an image captioning model as described in Subsection 3.1.

- **Concatenating CNN pooler output, with separated text models:** We conduct full fine-tuning on two distinct RoBERTa [8] models. The initial model, referred to as RoBERTa-Tweet (EXP6), is employed solely for tweet analysis to create a classifier for violent events. The second model, known as RoBERTa-Captions (EXP5), is utilized exclusively for image caption analysis. Additionally, an Inception-v3 [10] model is incorporated to process the accompanying images. The representation vector of the captions model (Embedding size of 768), the text model (Embedding size of 768), and the pooler output of the CNN (Vector size of 2048) are concatenated into a single vector that is then passed to a classification head consisting of a Multi-layer Perceptron [15] whose outputs are the class probabilities for each tweet (EXP7). The illustration for this proposal is described in Figure 2

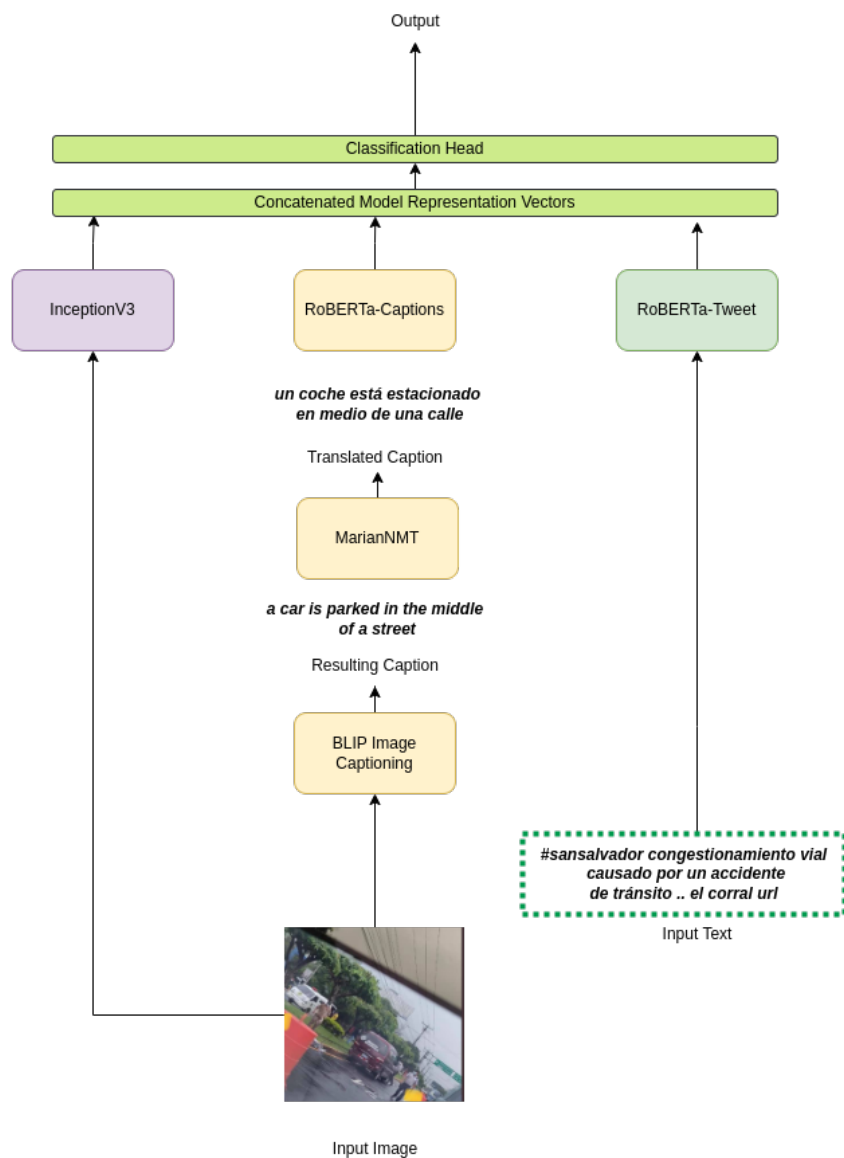


Figure 2: Illustration of the approach that concatenates features from the captions and text models with the CNN pooler output (See EXP7 Section 4.1.2).

- **Using Captions and Text in the same sentence to train a single model:** The second approach to merge the visual and the textual part is by connecting the captions and the tweet of the input data using a separator (`</s></s>` for RoBERTa model). This new representation is then passed to a single Transformer model and fine-tuned for each one of the two sub-tasks.

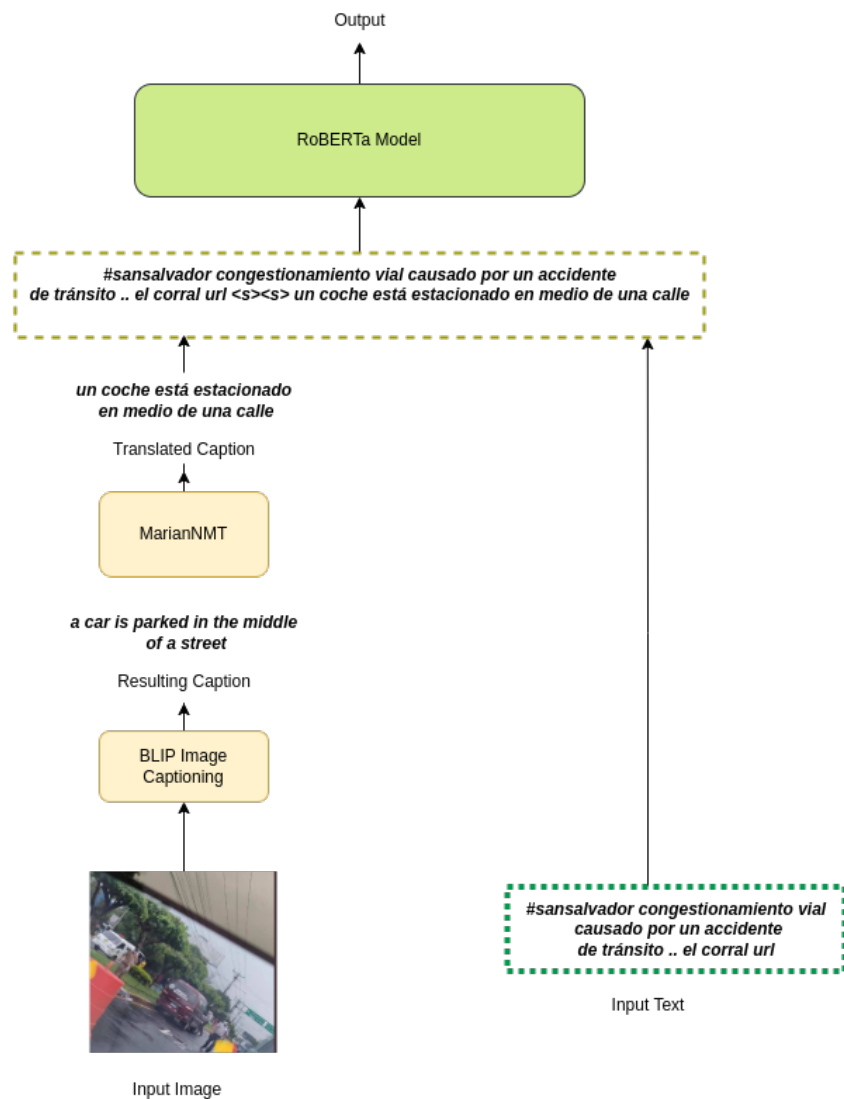


Figure 3: Illustration of the second approach that combines the caption and the text using a separator to pass it through the same Transformer model (See EXP8 Section 4.1.2.)

3.3. Proposed Randomized Data Oversampling, Weighted-Loss function and Ensemble Configuration.

Aside from the proposed-multi modal configurations, we believe that by tuning the parameters of the loss function and randomly increasing the number of examples of the under-represented categories we can increase the performance of the proposed models. In this work, we propose an approximation of the weights that may help to reduce the impact of the class imbalance. We propose to use data oversampling as well to make sure that at least one example of the

under-represented categories is contained in a training batch.

- Proposed Randomized Data Oversampling:** As described in Section 2, the dataset for Violent-Event-Categorization (VEC) has a high level of imbalance. Therefore we propose a randomized data oversampling strategy. To this aim, we consider the under-represented classes for *Murder* and *Robbery*. For each positive example $z \in \{Murder, Robbery\}$ we take a random number $x \sim \mathcal{U}(0, 1)$ and an oversampling factor f and replicate the number of copies of z by $x \cdot f$. This results in a higher number of positive examples of the under-represented classes. To avoid any bias, after performing this data oversampling, we shuffle the data randomly. This oversampling procedure is only applied to the training dataset.



Figure 4: Number of positive examples for each class of the training dataset before (LEFT) and after (RIGHT) the randomized data oversampling. From this figure, we see that after this oversampling procedure, the dataset has significantly more positive examples for classes *Murder* and *Robbery*.

- Weighted Loss-Function:** To reduce the impact of class imbalance for classes *Murder* and *Robbery* in VEC, we propose a modification of the Cross-Entropy Loss function adding weights to balance the importance of the under-represented classes. We use the expression of the Cross-Entropy Loss used by [16] that has the following formula.

$$\psi(x, y) = \{l_1, \dots, l_N, \} \text{ where}$$

$$l_n = -w_{y_n} \log \left(\frac{\exp(x_{n,y_n})}{\sum_{i=1}^N \exp(x_{i,y_i})} \right) \cdot \mathbb{I}_{y_n \neq \text{ignoreindex}}$$

Where N is the number of classes, x and y correspond to the input and the target respectively.

The weights w_n were adjusted according to the following formula.

$$w_n \approx \frac{1}{10P_c}$$

Where P_c is the probability of selecting a positive example of class c obtained by.

$$P_c = \frac{\sum_{s \in C} \mathbb{I}_{class(s)=c}}{|C|}$$

Where C represents the original dataset without data oversampling. From experimental results, the weights used in this work are presented in Table 2

Class	Associated Loss Weight
Traffic Accident	0.15
Murder	1.8
Robbery	1.7
Other	0.1

Table 2

Weights used to train the proposed model for Violent-Event categorization

- **Classification Boost using Ensembles:** As described in [17], model ensembles proved to be useful to boost the classification performance of the proposed models by stabilizing model predictions. To this aim, we train five models separately for each sub-task varying the weights initialization. Following up, we average the class probabilities obtained from the model outputs obtaining a final logit used for classification. This technique is applied to both sub-tasks to increase the model performance.

4. Results

In this section we present the results of the proposed models and proposed strategies in a custom dataset created from the original training dataset, these experiments helped us to determine the importance of parameter tuning to boost model performance and the best multi-modal approach for each one of the tasks to solve. In this section, we present the official results obtained for the VEI and VEC shared tasks.

4.1. Results in Custom Dataset

In order to evaluate the proposed models for sub-tasks 1 and 2, we created a stratified partition of the original training dataset consisting of 80% of total examples for training and the rest 20% for validation. We used all labeled examples available to train the final models for the final submission.

4.1.1. TEXT-ONLY experiments

To test the impact of parameter tuning on model performance we conducted an experiment to test how a Weighted Loss Function, Data Oversampling, and the combination of these two strategies may help to improve the classification capacities of a model. Especially when there is a high imbalance among classes like in sub-task 2. For these experiments, we set the oversampling factor f to 10. We tested the impact on the performance of the model in sub-task 2 by using a Weighted Loss Function (EXP1), a Randomized Data Oversampling (EXP2), and the combination of these two strategies (EXP3). To enable comparison, we conducted tests by training a model without utilizing any of the suggested strategies to enhance its performance (EXP0). For this comparison, we consider just the textual information and we use a RoBERTa [8] model with pre-trained weights from [14] as our base model. The learning rate for this experiment is set to $9e^{-6}$ with a weight decay of 0.1 and using AdamW [18] as model optimizer.

Boosting Strategy	Experiment ID	F1 Sub-task 2
None	EXP0	0.8453 ± 0.005
Weighted Cross-Entropy Loss (WCEL)	EXP1	0.881 ± 0.007
Randomize Data Oversampling (RDO)	EXP2	0.859 ± 0.001
RDO+WCEL	EXP3	0.889 ± 0.0182

Table 3

Boosting strategies used in this work to increase the model performance. The combination of the randomized data oversampling and the weighted loss function showed the highest increase in prediction performance.

From Table 3 we see that the combination of the Weighted Loss Function and the Data Oversampling increased the model capabilities (EXP3). It is worth mentioning that each one of the strategies increased significantly the classification performance being the Weighted-Loss-Function (EXP1) the boosting strategy with the highest increase on its own.

In this case, we see that by using these two boosting strategies together we obtained a better performance in Violent Event Categorization. While the enhancement in model performance may not be significantly superior compared to using any of the individual proposed strategies, it is evident that each strategy contributes to the model’s ability to identify important features in a distinct manner. The Weighted-Cross-Entropy approach assigns greater significance to underrepresented classes, whereas data oversampling ensures a higher representation of examples from these classes in the training batches.

As stated in [17], model ensembles may be useful to increase the classification performance of the models. In this case, we obtained a 0.01 – 0.02 increase when using ensembles compared to not using them. The inference time is not compromised as the ensemble procedure can be done in parallel.

4.1.2. Multi-modal configuration

The DA-VINCIS@IberLEF 2023 challenge [3] offers textual and visual data for each of the shared tasks, making it crucial to identify an appropriate multi-modal configuration that can effectively handle the VEC and VEI tasks. To obtain the best multi-modal configuration we obtained the

prediction performance of each one of the extracted features and the multi-modal approaches proposed in this work for both VEI and VEC. We consider the Macro Averaged F1 as our metric because is the same metric considered to evaluate the final submissions. All models were trained using data oversampling combined with a weighted loss function (EXP3). The models were trained using AdamW [18] as optimizer except for the Inception-v3 model where Adam [19] is used as optimizer. Table 4 shows the results from five runs with the different proposals.

Model	Experiment ID	F1 Sub-task 1	F1 Sub-task 2
Inception-V3[10]	EXP4	0.8029 ± 0.006	0.45 ± 0.03
RoBERTa-Captions	EXP5	0.82 ± 0.02	0.512 ± 0.0131
RoBERTa-Tweet	EXP6	0.936 ± 0.005	0.889 ± 0.0182
Outputs concatenation	EXP7	0.919 ± 0.002	0.842 ± 0.0107
Joining Caption and Tweet with separator	EXP8	0.943 ± 0.003	0.892 ± 0.0072

Table 4

Macro average F1 obtained for the different proposals. We see that the model using the captions and the tweet joint by a separator performs better than the other evaluated proposals.

From Table 4 we see that the features obtained from the images do not provide sufficient information to either detect or categorize violent events (EXP4). However, when this information is combined with the tweet using a separator as in our proposal (EXP8), it slightly increases the classification performance of the text model. This may be due to the inherent Self-attention mechanism of all the Transformer models. Self-attention is an attention mechanism that establishes connections between various positions within a single sequence, enabling the computation of a representation for the entire sequence [13]. Therefore a Transformer model is able to determine which information (from text and image) is important to create the representation vector of the sequence to correctly identify or categorize a violent event.

To demonstrate this, we employ an integrated gradients approach to extract significant features from the texts and ascertain the words that are pertinent for classifying the presence of a violent event.

Figure 5 reveals that certain tokens within the caption have an influence on the model to infer a positive prediction, indicating its capability to identify the crucial features within the combined input text and generated caption. Based on the observations from Figure 5, it becomes apparent that words such as "accidente" (accident) and "carretera" (road) play a crucial role in the model's ability to recognize a Violent Event mentioned in the tweet. Notably, some of these important words are included in the generated caption. Figure 6 illustrates a tweet in which no violent events were detected. It is noteworthy that words like "emoji" play a role in the model's negative prediction. Similarly to the positive example, some of the significant words utilized to classify the tweet as a "Non Violent" example are present within the caption generated by the associated image.

4.2. Official submissions to the Violent-Event-Detection shared task (Sub-task 1)

We did two submissions for sub-task 1, the first one denoted by RUN1 corresponds to an ensemble of five models with different weight initialization, and the second submission (RUN2) corresponds

DaVinci's results		
Run	Model	F1 Score
RUN1	Model Ensemble	0.92638
RUN2	Single Model	0.92418
<i>Second Best Team</i>	-	0.9203
<i>Baseline</i>	-	0.8948
<i>Average Performance</i>	-	0.8911

Table 5

Official results for sub-task 1. We see that our ensemble proposal achieved the best result for sub-task 1. However, the single model approach also has a higher F1 score than the second-best model submitted.

DaVinci's results		
Run	Model	F1 Score
RUN1	Model Ensemble	0.842074
RUN2	Single Model	0.83038
<i>First Best Team</i>	-	0.8797
<i>Second Best Team</i>	-	0.8733
<i>Baseline</i>	-	0.8427
<i>Average Performance</i>	-	0.8071

Table 6

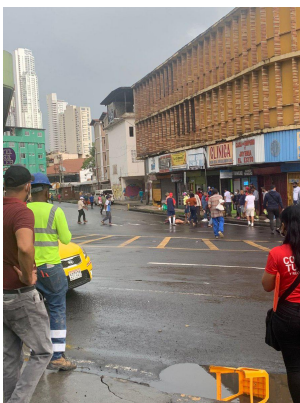
Official results for sub-task 2. In this case, the ensemble boosting strategy showed a higher increase in model performance compared to the single model.

diversity. This approach enables the model to learn broader patterns, thereby improving the categorization of Violent Events. Similarly, in the case of Weighted-Cross-Entropy, evolutionary strategies like the Covariance Matrix Adaptation (CMA-ES) [21] algorithm can be utilized to determine optimal weights that maximize the F1 score for the under-represented classes.

4.3. Error Analysis

According to the results obtained from the predictions made by the proposed model, the classes *Murder* and *Robbery* have the lowest F1 score of all classes in the violent event categorization. In this section, we present some misclassified examples of these two classes. The first example corresponds to an example of the *Murder* class that was misclassified and assigned to the *Other* class.

As we see from the picture associated with the text, it does not provide any useful information that could be related to a murder. The text contains words like *investigación* (investigation) suggesting that the content described in the tweet is not a confirmed fact.



el @usuario inició investigación por delito de homicidio, luego de un hecho registrado en la 5 de mayo, donde fallecieron dos personas 'en el

*área se recopilan indicios' señala la entidad
#exitosanoticias url </s></s> un grupo de*

personas de pie alrededor de una calle

The information shown below is labeled as a positive example of *Robbery*, however, the model predicted this example as *Murder*. It is suggested that the word *disparo* (shot) is mostly associated by the model with murder. The image related to the text does not provide a lot of information similar to the previous example. This shows that the visual part lacks information to complement the textual information.



*naucalpan edomex emoji coche de policía emoji
emoji ambulancia emoji una mujer recibió un
disparo durante un asalto a bordo de una combi
los agresores huyeron fue en la esquina de
avenida naucalpan y calle allende, en la colonia
hidalgo @usuario url </s></s> un hombre está
durmiendo en el suelo en un autobús*

5. Conclusions

This paper describes our participation at the DA-VINCIS@IberLEF 2023 challenge on the *Violent-Event-Identification* and *Violent-Event-Categorization* sub-tasks. Our approaches rely on optimal parameter tuning and multi-modal strategies to solve both sub-tasks. The proposed solutions showed a great performance in the VEI task obtaining the best results in the challenge. For the Violent-Event-Categorization task, further parameter tuning is required to increase model results. Based on the experimental findings presented in this research paper, it is evident that achieving better model performance necessitates greater diversity in text generation. Consequently, we propose the incorporation of data augmentation strategies, such as back-translation or generative models, to generate data specifically for the underrepresented classes. In order to obtain more precise weights for the cross-entropy loss, we hypothesize that employing evolutionary strategies will lead to further improvements in model performance.

Acknowledgments

The authors thank CONACYT, INAOE and CIMAT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies (*Laboratorio de Supercómputo: Plataforma de Aprendizaje Profundo*) with the project "*Identification of Aggressive and Offensive text through specialized BERT's ensembles*" and CIMAT Bajío Supercomputing Laboratory (#300832). Esaú Villatoro-Tello, was supported by Idiap Research Institute during the elaboration of this work.

References

- [1] K. C. Miller, J. L. Nelson, “dark participation” without representation: A structural approach to journalism’s social media crisis, *Social Media+ Society* 8 (2022) 20563051221129156.
- [2] A. S. Veenstra, N. Iyer, C. S. Park, F. Alajmi, Twitter as “a journalistic substitute”? examining #wionion tweeters’ behavior and self-perception, *Journalism* 16 (2015) 488–504.
- [3] H. Jarquín-Vásquez, D. I. H. Fariás, J. Arellano, H. J. Escalante, L. V. nor Pineda, M. M. 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).
- [4] L. J. Arellano, H. J. Escalante, L. Villaseñor Pineda, M. Montes y Gómez, F. Sanchez-Vega, Overview of da-vincis at iberlef 2022: Detection of aggressive and violent incidents from social media in spanish (2022).
- [5] D. Vallejo-Aldana, A. P. López-Monroy, E. Villatoro-Tello, 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, 2022.
- [6] G. Qin, J. He, Q. Bai, N. Lin, J. Wang, K. Zhou, D. Zhou, A. Yang, 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, 2022.
- [7] P. Turón, N. Perez, A. Garcia-Pablos, E. Zotova, M. Cuadros, 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, 2022.
- [8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).
- [9] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: *International Conference on Machine Learning*, PMLR, 2022, pp. 12888–12900.
- [10] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
- [11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.
- [12] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, et al., Marian: Fast neural machine translation in c++, *arXiv preprint arXiv:1804.00344* (2018).
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [14] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, Robertuito: a pre-trained language model for social media text in spanish, *arXiv preprint arXiv:2111.09453* (2021).
- [15] F. Rosenblatt, Principles of neurodynamics. perceptrons and the theory of brain mecha-

- nisms, Technical Report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.
- [16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, *Advances in neural information processing systems* 32 (2019).
 - [17] J. Briskilal, C. Subalalitha, An ensemble model for classifying idioms and literal texts using bert and roberta, *Information Processing & Management* 59 (2022) 102756.
 - [18] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, *arXiv preprint arXiv:1711.05101* (2017).
 - [19] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).
 - [20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.
 - [21] N. Hansen, S. D. Müller, P. Koumoutsakos, Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es), *Evolutionary computation* 11 (2003) 1–18.