

# SINAI Participation at DA-VINCIS Task in IberLEF 2023: Data Augmentation for Multimodal Classification

Alberto José Gutiérrez-Megías<sup>1</sup>, Sergiu Stoia<sup>1</sup>, Fernando Martínez-Santiago<sup>1</sup>, Luis Alfonso Ureña-López<sup>1</sup> and Arturo Montejo-Ráez<sup>1,†</sup>

<sup>2</sup>University of Jaén, Las Lagunillas s/n, Jaén, 23071, Spain

## Abstract

This paper describes the participation of the SINAI team in the Detection of Aggressive and Violent INCIDENTS from Social Media in Spanish (DA-VINCIS) task organized at the evaluation campaign IberLEF 2023. The system combines the encoding of text and images with separated pretrained neural networks. The network was fine-tuned on an augmented dataset from the provided training data. The proposed approach obtained, in task 1, an F1 score of 0.9165 and, in task 2, the second-best result, with an F1 score of 0.8733. The results suggest that data augmentation and the fusion of modality-specialized encoders are valid strategies for achieving state-of-the-art results for the automatic classification of data that combine visual and textual information.

## Keywords

Violence Detection, IberLEF, DA-VINCIS, Multimodal, Social Media, Natural Language Processing, Spanish

## 1. Introduction

Although chronic violence builds on historical legacies of social and political violence, oppression, exclusion, and armed conflict, it is also molded by contemporary dynamics such as rapidly evolving forms of governance, information technologies, climate change, the intensified dynamics of globalization, and other factors [1].

Violent comments flood social networks, this new form of communication is easily accessible to most of the world's population. Social media platforms have attempted to curb the presence of users who engage in violent comments by implementing manual tools. However, the responsibility of identifying and silencing such comments lies with the platform's users themselves. Recognizing indications of violence within a corpus of text, especially in the absence of contextual information, is a challenging undertaking.

---

*IberLEF, September 2023, Jaén, Spain*


\*Corresponding author.

✉ agmegias@ujaen.es (A. J. Gutiérrez-Megías); sstoia@ujaen.es (S. Stoia); dofer@ujaen.es (F. Martínez-Santiago); laurena@ujaen.es (L. A. Ureña-López); amontejo@ujaen.es (A. Montejo-Ráez)

🆔 0009-0006-0834-8177 (A. J. Gutiérrez-Megías); 0009-0009-2599-2820 (S. Stoia); 0000-0002-1480-1752 (F. Martínez-Santiago); 0000-0001-7540-4059 (L. A. Ureña-López); 0000-0002-8643-2714 (A. Montejo-Ráez)



© 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)

In the tasks proposed by DA-VINCIS lab [2] at the IberLEF2023 evaluation forum [3], a corpus with textual information is supported by images, with the aim of improving the automatic detection of violence on Spanish tweets.

DA-VINCIS task is divided into two subtasks: (1) violent event identification, which consists in the classification of tweets as violent or non-violent (binary classification); (2) violent event category recognition, which is a multi-label task to identify among an *accident*, *murder*, *robbery*, and *other*, where several labels could be selected in the same comment.

In this paper, we present a multimodal proposal, where the image and the text information are encoded by pretrained models. We propose two pretrained models, which have been fine-tuned, to encode both visual and textual information. Resulting tensors are then concatenated before passing to a feed forward network for final label prediction. To train the models, we have followed a data augmentation strategy.

The paper is organized as follows: in Section 2 a brief review of related works on methods and ideas for the detection of violence in text and multimodal systems is given. In Section 3, an explanation of the corpus and the preprocessing followed to obtain more data through augmentation for training of the model is provided. Then, in Section 4, the architecture of the proposed model is explained in more detail, describing the pretrained models used and detailing the strategy to search for hyperparameters, as well as the use of weights to deal with unbalanced classes in the dataset, as in the case of the Violent event category recognition task. In Section 5 we will show the experiments performed and the results of the best models found. Finally, in Section 6 we will draw some conclusions from the results obtained and discuss possible future work.

## 2. Related work

The widespread access to social networks in recent years has demanded detecting and removing sensitive content to prevent minors from viewing it. This content can range from accidents, aggressions, personal insults, sensitive news and alike. Early exposure to violent content on the internet is related to desensitization to violence [4].

Types of violence such as gender-based violence are an under-reported public health problem, and non-physical forms are recognized as psychological violence. This type of violence can sometimes be detected through social networks. Universal Sentence Encoder (USE) has been used in conjunction with multiple machine learning algorithms for the task of violence detection [5].

The detection of violence is a novel task, most of the classifications that we can find these years are classifications by text or images. Probabilistic models such as Latent Dirichlet Allocation (LDA) topic models [6] have been used for the classification of harmful content on websites. The recognition of violence contained in images, a multi-view maximum entropy discriminant model for learning with different numbers of views has been proposed [7].

With the introduction of Transformers [8] and pretrained models, classification results can be improved and optimized, using models such as RoBERTa for short-text classification problems [9]. There is already work on the classification of violent texts, focusing on intimate partner violence using models such as BERT and RoBERTa [10]. Information about the same

phenomenon can be acquired from different types of detectors.

We use the term "modality" for each of these acquisition frameworks. It is rare that a single modality provides complete knowledge of the phenomenon of interest. The increasing availability of several modalities reporting on the same system introduces new degrees of freedom [11]. Multimodal models are systems designed to provide the capacity of working with different types of inputs to train or predict tasks. Multimodal models have been used to fuse textual information extracted by Optical Character Recognition (OCR) with visual Convolutional Neural Network (CNN) methods [12].

In this paper, we will use the concatenation of textual and visual information provided by images associated with a Twitter comment in Spanish in order to classify it as violent or not.

### 3. Data Description and Preprocessing

The provided training data comprises three text files containing comments sourced from Twitter, a folder containing images, and corresponding labels for each task. Each subtask uses the same training data, with the labels being the main difference. The relationship between images and text is closely intertwined, with each tweet item accompanied by at least one image, and a maximum of four. We have 2,996 training entries, each associated with a label according to the subtask. The dataset contains a cumulative count of 4,267 images, spread across these entries. On average, each text has 1.42 images associated with it, since the median of the data is one image per entry, we can conclude that most of the texts have only one image associated with them, with 4 images being the maximum for each text. Table 2 shows the classification of the labels for each data. The first subtask exhibits a reasonable balance among its classes. However, in the case of the multi-label task, a notable imbalance exists within the *murder* and *theft* classes.

Labels of subtask 1 contain two values, (0) for non-violent entries and (1) for violent entries. For the second subtask, it is provided a vector with four possible values (*accident*, *murder*, *robbery*, *other*). These values are not exclusive of each other, the same entry can be a robbery and a murder at the same time, so we talk about a multi-label problem.

Text entries were processed as follows:

- Replacing links with the tag "<URL>".
- Replacing users with the tag "<USER>".
- Removing emojis.

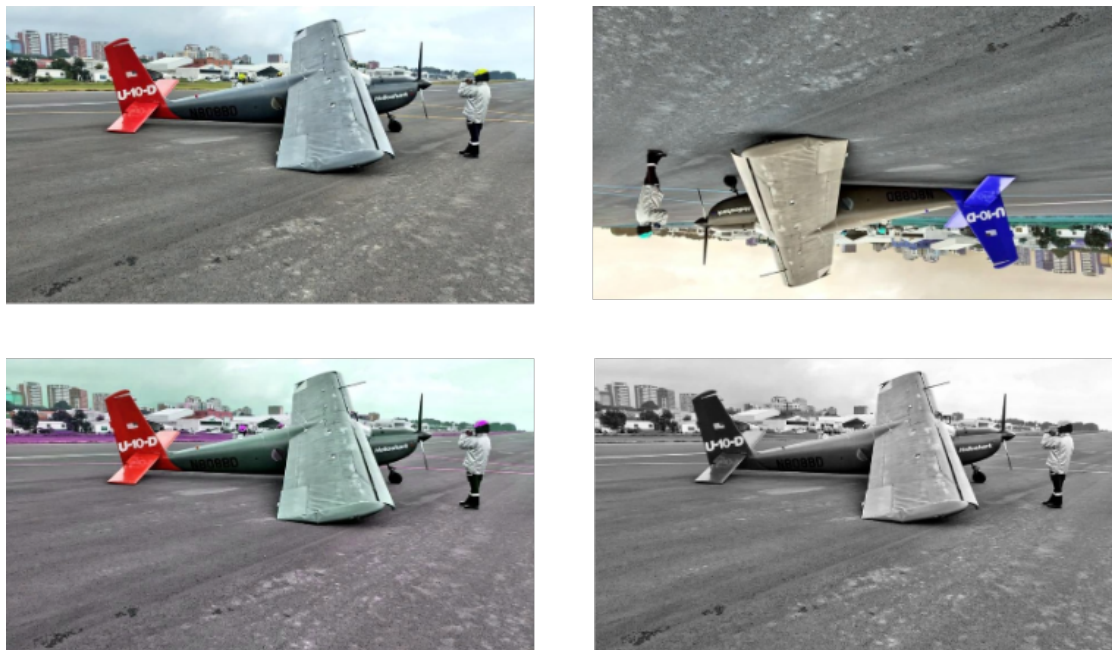
Due to the limited training data, we have applied various data augmentation techniques. These include modifications to the original images using image modification libraries, and back-translation for the text. The details are given in the next section.

#### 3.1. Data Augmentation

The first step for the data augmentation has been to split a tweet into several entries, the same tweet has been multiplied by the number of images associated to it, i.e. [*image 1*, *image 2*, *image 3*], *text*; would be the original entry. After splitting, each sample would have one entry for each image.

Following the previous step, it is observed that multiple images will be linked to the same text entry. To address this situation, a back-translation approach will be employed for texts that are identical within an entry but have distinct associated images. For instance, if a particular entry has repeated the same text three times, the last text in the sequence will undergo translation from Spanish to English, and vice versa, three times. This process introduces minor variations to each text within the same tweet, thereby diversifying the dataset.

Finally, each image has been used to generate three different versions by applying three different transformations, using the *albumentations* [13] library. The initial transformation involved applying a horizontal flip followed by a vertical flip to the image, with a 50% probability of applying this transformation. The second transformation involves converting the image to monochrome, resulting in a grayscale representation. Additionally, a blur effect is applied to the image, with a 40% probability of this transformation being implemented. These transformations can be seen in figure 1, and the result of the training dataset in Table 1 below.



**Figure 1:** Image transformations used for data augmentation.

The comparison between the original data and those obtained after data augmentation to improve the quality of the training set can be seen in Table 2. There is a class imbalance in subtask 2, where there are significantly fewer values for *murder* and *robbery*, compared to the other two labels. This negatively affects multi-label training. One solution is to use weights for each class. A more in-depth explanation of the weighting scheme followed for dealing with class imbalance is given in Section ??.

**Table 1**

Examples from the augmented dataset.

Imagen	Text
Image 1 original	Morale #EEUU sufrió una derrota vergonzosa ante ...
Image 1 modification 1	Morale #USA sufrió una vergonzosa derrota contra ...
Image 1 modification 2	Morale #USA sufrió una vergonzosa derrota contra ...
Image 1 modification 3	Morale #USA sufrió una vergonzosa derrota contra ...
Image 2 original	Tus acciones te hacen ser una bella persona ...
Image 2 modification 1	Tus acciones te hacen una persona hermosa, ...
...	

**Table 2**

Difference in data distribution of both tasks by labels between the original and the augmented data.

Dataset	Subtask 1		Subtask 2			
	Violence	No-violence	Accident	Murder	Robbery	Other
Original data	1277	1719	940	182	190	1719
Augmented data	7579	9480	5748	927	1112	9480

## 4. Our approach

We can split the process followed by our multimodal approach into the following steps:

1. Before training, images and texts are preprocessed separately, through tokenizers and truncating the sequences before being used as inputs to the model. The output obtained from these tokenizers will be the input together with the labels associated with each piece of data for the model.
2. The architecture is based on the use of two pretrained models that tokenize images and text separately. The results obtained are stored separately to be used as input to the pretrained models.
3. Within the multimodal model, the features obtained by the tokenizers are passed as input for each pretrained model depending on whether they are textual features or images.
4. The output tensor of each of these models is concatenated as shown in Figure 2.

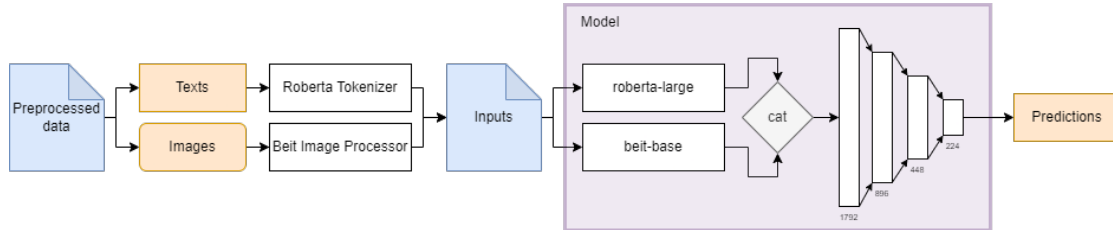
### 4.1. Pretrained models

Two pretrained models are used in the system architecture. The text processing task is based on the model is RoBERTa (A Robustly Optimised BERT Pretraining Approach) [14] and BEiT (BERT pretraining if Image Transformers) [15] for visual features extraction:

- **RoBERTa Large:** RoBERTa is a self-monitored pretrained transformer model. This means that it is pretrained with the Masked language modeling (MLM) objective only with the raw texts without human supervision. It should be acknowledged that, inadvertently, the English variant of RoBERTa (the original version) was utilized instead of a version explicitly or implicitly tailored to support the Spanish language.

The use of RoBERTa Large implies the model must adapt to Spanish text, in which case performance is affected than if we had used a RoBERTa model in its Spanish version. Remarkably, the final system demonstrated satisfactory performance, indicating that RoBERTa exhibited proficiency in handling Spanish or could be effectively fine-tuned using Spanish text corpora.

- **BEiT Base patch16-244:** This model is a pretrained BEiT model. The BEiT model is a Vision Transformer (ViT), which is a transformer encoder model (BERT-like). The model learns an internal representation of the images that can then be used to extract useful features. In addition, if you have tagged images, it can classify them.



**Figure 2:** The system architecture of the multimodal model and the processing of input data.

## 4.2. Architecture

The system architecture starts with the preprocessed data. Images and text are processed independently: texts by Roberta Tokenizer to obtain the inputs for the transformer model. Images go through BEiT Image Processor to obtain *pixel values*. These will be the inputs to the visual network.

The outputs of these models will be concatenated into a single tensor, which will be the input of the classification layers. This final feed forward network (the classification head) consists of four layers, each with a ReLu activation layer and a Dropout layer with a probability in the range of [0.4, 0.5], depending on the hyperparameter optimization. Each of the layers has a different size, decreasing by half, until reaching the last layer, to force final encoding. The size of the output layer depends on the type of subtask being addressed. For binary classification, the output size would be 1, while for multi-label classification, it would be 4. For both tasks, a final sigmoid function is applied. A threshold of 0.5 over that final output will serve to solve both, binary and multi-label classification.

The choice of loss function differs based on the specific subtask the model is trained on. For subtask 1, we used binary cross entropy without weight modification because both classes (*violence, non-violence*) are well balanced, as seen in Table 2. For subtask 2, on the other hand, the data is unbalanced. Therefore, in the cross entropy loss function, a weight for positive samples is applied. This particular loss function can be customized by providing a vector of weights for the positive classes. Considering subtask 2, it is necessary to utilize weights, with one weight assigned to each of the classes. For a given class, positive counts are the number of samples assigned to that label and the negative count the other remaining labels (see Equation 1). Therefore, the more samples are found for a class, the lower the loss calibration weight.

$$Weight = \frac{N_{neg}}{N_{pos}} \quad (1)$$

The weights used to train the multi-label task for each label are shown in Table 3, taking into consideration the original corpus provided by the competition to calculate the weights.

**Table 3**

Class weights used for training in subtask 2.

Label	# positive	# negative	Weight
Accident	940	2091	2.22
Murder	182	2849	15.65
Robbery	190	2841	14.95
Other	1312	1719	0.76

### 4.3. Hyperparameters optimization

During the optimization process, various ranges of values for four hyperparameters were explored following a grid search approach, as illustrated in Table 4. To identify the best values, an early stopping technique was employed, where the stopping condition was evaluated during the training process using a development split. This condition is that the macro metric F1 stops improving during three consecutive epochs. This way, the number of epochs to be used for a model with specific parameters can be reliably adjusted.

**Table 4**

Ranges of hyperparameter values.

Hyperparameter	Values
Epochs	[1, 20]
Batch size	[8, 16]
Learning rate	[5e-6, 1e-4]
Dropout	[0.4, 0.5]

After a series of experiments, using 80% of the data as training data, and the remaining 20% of data as validation data, the best hyperparameters found for each of the subtasks according to the results obtained from validations are represented in Table 5 for subtask 1, and Table 6 for subtask 2 in Section 5.

## 5. Experiments and results

After data augmentation, the dataset has 17,059 samples and 16 entries referring to the same text could be found in the dataset. The text might not be similar to the original due to the back-translation technique. To maintain the integrity of the dataset and prevent similar texts from being split between the test and validation sets, a manual split was performed. The goal was to ensure that all similar texts, which have undergone the back-translation process, remain within the same dataset. The split for the local experiments was 80% for training and 20%

for validation. The selected validation set does not include the modifications made for data augmentation to be faithful to the real data to be finally predicted in the final phase of the competition.

After identifying the best-performing models, they were trained using the entire dataset. With these trained models, predictions were made for the test samples, and the resulting labels were submitted to the Codalab platform, as instructed by the organizers.

In Tables 5 and 6, the development dataset refers to the local validations using 80% of the data as training and the remaining 20% as validation as explained in previous sections. The validation dataset was provided by the competition for testing. The best F1-score result has been obtained with that dataset because the training set for the validation ranking is richer than with the local experiments. The best results obtained and the models' parameters are shown in Tables 5 and 6.

**Table 5**

Best results with the validation dataset for violent event identification (subtask 1).

Dataset	Splitting	Learning rate	Epochs	Batch size	F1-score	Precision	Recall
Development	8:2	1e-05	4	16	0.9214	<b>0.9218</b>	0.9211
Development	8:2	5e-05	3	8	0.8105	0.8137	0.8101
Validation	full	1e-05	4	16	<b>0.9231</b>	0.9098	<b>0.9367</b>

**Table 6**

Best results with the validation dataset for violent event category recognition (subtask 2).

Dataset	Splitting	Learning rate	Epochs	Batch size	F1-score	Precision	Recall
Development	8:2	1e-05	4	16	0.8565	<b>0.8971</b>	0.8293
Development	8:2	5e-06	7	16	0.8411	0.8015	0.8917
Validation	full	5e-06	7	16	<b>0.8976</b>	0.8796	<b>0.9184</b>

The models used for the final test of the competition have been those that have given the best results in the official validation dataset, shown in Tables 5 and 6 above.

In the competition's final results for the violent event identification task (subtask 1) we obtained an F1-score of 0.9165, giving us the fourth-best score. Table 7 shows the best results of the teams in the leaderboard and their scores for this task.

**Table 7**

Official subtask 1 ranking results.

Team	F1-score	Precision	Recall
danielvallejo237	0.9264	0.9302	0.9226
EstebanPonce	0.9203	0.9006	0.9409
Jorge	0.9186	0.9067	0.9308
<b>agmegias</b>	<b>0.9165</b>	<b>0.8951</b>	<b>0.9389</b>
csuazob	0.9100	0.8939	0.9267
Arnold	0.9069	0.9014	0.9124

Finally, for the violent event category recognition task (subtask 2) we obtained an F1-score of



0.8733, obtaining the second-best score. Table 8 shows the best results for this task, as in the previous table.

**Table 8**

Official subtask 2 ranking results.

Team	F1-score	Precision	Recall
EstebanPonce	0.8797	0.8737	0.8864
<b>agmegias</b>	<b>0.8733</b>	<b>0.8523</b>	<b>0.8973</b>
Jorge	0.8698	0.8622	0.8784
Arnold	0.8492	0.8305	0.8715
csuazob	0.8490	0.8441	0.8577
HoracioJarquin	0.8427	0.7663	0.9407

All systems have been launched on NVIDIA A100-SXM4-40GB, with a maximum of 20 epochs with early stopping looking for the best F1-score result, as explained in Section 4.3. The training of the final models has taken approximately 40~60 minutes for each task.

### 5.1. Voting system for final prediction

The results obtained from the predictions still need to be processed. As explained above, the data with which the classifier network is trained has the following structure: [image | text].

For each tweet, if the original data has three images, three different predictions will be generated. A voting system has been applied to decide which predictive result is the selected one for the tweet. If for the same tweet, there are more or the same positive tags (evidence of violence) than negative tags, it will be labeled as violent. In the case of subtask 2, for each of the options (*accident*, *robbery*, *murder*, *other*), a vote is made for each of them, following the same voting rule as in subtask 1 explained above.

## 6. Conclusions and future work

In this paper, we have developed a multimodal system capable of processing textual and visual information by concatenating them in order to solve classification and multi-label classification problems over the DA-VINCIS lab at IberLEF2023, which provided a corpus of tweets with Spanish images to solve each task it presents.

We have achieved 4th place in the violence classification task and 2nd place in the violence category classification task. The results are promising, and in the future, we will develop other ideas such as exploring further augmentation techniques or using different types of pretrained models for linking and processing textual and visual information. Actually, regarding textual information, we plan to replicate experiments using models supporting Spanish, like XLM-RoBERTa [16], mDeBERTa [17] or MarIA’s project ones [18]. Also, soft voting instead of hard one applied will be explored. Besides, alternatives to a simple concatenation of features before the classification head may not be the best approach for feature combinations. In this sense, we will study other possibilities, like gating or attention, to generate multimodal encoding.

## Acknowledgments

This work has been partially supported by WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government, and projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government, and project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government.

## References

- [1] T. M. Adams, How chronic violence affects human development, social relations, and the practice of citizenship: A systemic framework for action, Washington, DC: Woodrow Wilson International Center for Scholars (2017).
- [2] H. Jarquín-Vásquez, D. I. H. Farías, 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).
- [3] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, *Procesamiento del Lenguaje Natural* 71 (2023).
- [4] E. Kennedy, Early Online Graphic Content Exposure and the Development of Desensitisation to Violence, Ph.D. thesis, Dublin, National College of Ireland, 2023.
- [5] P. Trinh Ha, R. D'Silva, E. Chen, M. Koyutürk, G. Karakurt, Identification of intimate partner violence from free text descriptions in social media, *Journal of Computational Social Science* 5 (2022) 1207–1233.
- [6] S. Liu, T. Forss, New classification models for detecting hate and violence web content, in: 2015 7th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3K), volume 1, IEEE, 2015, pp. 487–495.
- [7] S. Sun, Y. Liu, L. Mao, Multi-view learning for visual violence recognition with maximum entropy discrimination and deep features, *Information Fusion* 50 (2019) 43–53.
- [8] 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).
- [9] Z. Guo, L. Zhu, L. Han, Research on short text classification based on roberta-textrcnn, in: 2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI), 2021, pp. 845–849. doi:10.1109/CISAI54367.2021.00171.
- [10] M. A. Al-Garadi, S. Kim, Y. Guo, E. Warren, Y.-C. Yang, S. Lakamana, A. Sarker, Natural language model for automatic identification of intimate partner violence reports from twitter, *Array* 15 (2022) 100217.
- [11] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: An overview of methods, challenges, and prospects, *Proceedings of the IEEE* 103 (2015) 1449–1477. doi:10.1109/JPROC.2015.2460697.
- [12] R. Jain, C. Wigington, Multimodal document image classification, in: 2019 International

- Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 71–77. doi:10.1109/ICDAR.2019.00021.
- [13] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A. A. Kalinin, Albumentations: Fast and flexible image augmentations, *Information* 11 (2020). URL: <https://www.mdpi.com/2078-2489/11/2/125>. doi:10.3390/info11020125.
- [14] 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, *CoRR abs/1907.11692* (2019). URL: <http://arxiv.org/abs/1907.11692>. arXiv:1907.11692.
- [15] H. Bao, L. Dong, S. Piao, F. Wei, Beit: Bert pre-training of image transformers, *arXiv preprint arXiv:2106.08254* (2021).
- [16] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, *CoRR abs/1911.02116* (2019). URL: <http://arxiv.org/abs/1911.02116>. arXiv:1911.02116.
- [17] P. He, X. Liu, J. Gao, W. Chen, DEBERTA: Decoding-Enhanced BERT with Disentangled Attention, in: *International Conference on Learning Representations*, 2021. URL: <https://openreview.net/forum?id=XPZlaotutsD>.
- [18] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, *Procesamiento del Lenguaje Natural* 68 (2022). URL: <https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley>. doi:10.26342/2022-68-3.