

ITAINNOVA@DA-VINCIS: A Tale of Transformers and Simple Optimization Techniques

Rosa María Montañés-Salas^{1,*}, Rafael del-Hoyo-Alonso¹ and Paula Peña-Larena¹

¹*Technological Institute of Aragón (ITAINNOVA), María de Luna, 7–8, Zaragoza, Spain*

Abstract

This paper describes the participation of ITAINNOVA at the “Detection of Aggressive and Violent Incidents from Social Media in Spanish” task (DA-VINCIS), framed within the evaluation forum IberLEF (Iberian Languages Evaluation Forum), a shared evaluation campaign for Natural Language Processing (NLP) systems in Spanish and other Iberian languages. This work explores the integration of state of the art transformer-based language models with conventional machine learning optimization techniques as well as Natural Language Processing (NLP) standard methods, with the objective of identifying and categorising tweets in Spanish concerning to violent events. The system implemented applies the backtranslation text augmentation method and thereafter employs a hyperparameter tuning strategy to improve finetuned language models performance. Finally, a voting ensemble approach is employed on top of each model predictions. The experimental results obtained are quite encouraging, since a 0.76 F1 score has been reached at binary classification and a 0.50 F1 on the multilabel subtask.

Keywords

NLP, Transformers, Optimization techniques, Social Networks

1. Introduction

The “Detection of Aggressive and Violent Incidents from Social Media in Spanish” (DA-VINCIS) task [1] is one of the challenging research problems related to harmful content proposed within the IberLEF (Iberian Languages Evaluation Forum) campaign, which is celebrated under the umbrella of the International Conference of the Spanish Society for Natural Language Processing (SEPLN). The aim of the proposed task is to promote the current state of development of the detection and monitoring of aggressive and violent events disseminated through social media networks in Spanish. Social networks such as Twitter are excellent means to publicly share news, opinions and events by any user in a fast and concise way, so that a critical or violent incident can reach other users and relevant authorities rapidly, thus minimizing its negative consequences. In order to face that problem, DA-VINCIS feature two subtasks: the first one is “Violent event identification” which aims to determine if a tweet express an aggressive or violent incident as a binary text classification problem; the second one “Violent category recognition”, is meant to recognize which kind of incident is mentioned in the text as multi-class


IberLEF2022, September 2022, A Coruña, Spain.

*Corresponding author.

✉ rmontanes@itainnova.es (R. M. Montañés-Salas); rdelhoyo@itainnova.es (R. del-Hoyo-Alonso); ppena@itainnova.es (P. Peña-Larena)



© 2022 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)

multi-label classification problem. The categories to be identified are: Accident, Homicide, Non-Violent-incident, Robbery, Kidnapping.

The team from ITAINNOVA propose an approach based on the state of the art Natural Language Processing (NLP) algorithms, that is, the Transformer-based Language Models (LMs) capable of performing a wide variety of tasks through learning highly abstract language concepts and representations. These deep learning architectures are pre-trained on a large corpus with general learning objectives, such as filling masked tokens or interrelating sentences, and then fine-tuned on downstream tasks such as text categorization. We have enriched those models potential by applying different machine learning optimization and text processing techniques getting to improve the overall performance, and consequently obtaining quite satisfactory results on the identification and classification of violent incidents in tweets in Spanish.

The paper is structured as follows: after the introduction, a set of works which has inspired our approaches are described. In section 3 the system description is presented, followed by the details of the experiments carried out in section 4. Results of those experiments are presented as well. Finally, in section 5 a summary of the main conclusions drawn during the experimentation and future working directions are exposed.

2. Related work

Every day since the arrival of the deep learning models known as Transformers [2], we see new models in the news, with a larger number of connections and parameters, GPT1-3, GATO, etc.[3][4]. Models capable of obtaining the human-like results in most of the classic problems of Natural Language Processing, such as entity recognition (NER) [5], recommendation [6], opinion classification [7], or more advanced techniques such as paraphrasing or generation of summaries [8]. Initially, it was revealed that the number of parameters and the model's competence in certain tests had a relationship. The performance and problem-solving ability are correlated with the number of model parameters [9]. Recent research has revealed that the quality and size of the corpus, as well as the number of parameters, are critical to the results obtained [10]. Many layers and neurons are underused in these large language models, so smaller or more simplified models with the same efficiency can be found. BERT [11] and Roberta [12], as well as others [13], are examples of model simplification techniques used for this purpose. These techniques allow the reduction of initial models size while maintaining efficiency, allowing faster processing.

Harmful content detection in text, particularly in social media platforms [14], is a challenge where new transformers models can get greater outcomes, but there aren't a lot of texts conveniently annotated to train the models on. Data augmentation refers to techniques for increasing the quantity of data available by combining slightly modified copies of current data with freshly produced synthetic data. Text augmentation, is the same approach in texts for generating training examples automatically. As a result, in this scenario, it can assist in improving the performance of Deep Learning Models. Recently, several of text enhancement techniques have been presented. [15] and [16] and they are methods to improve classification tasks when corpus are smaller like in this case.

The lack of models in non-English languages is the biggest issue in this language modelling

race. The majority are in English or multilingual, with the latter having a lower ability to solve problems in languages like Spanish. We are beginning to have models in Spanish, trained to a sufficient capacity, thanks to the National Language Plan and the BSC's (Barcelona Supercomputing Center) initiative [17]¹. We're still a long way from having great Spanish models, but this kind of initiative is to be applauded.

Despite the great advances achieved with transformer-based language models, there is still room to apply more traditional methods such as model ensembling, which in these cases help to enhance the results obtained in complex human language comprehension tasks [18][19].

3. System description

ITAINNOVA's system is designed and implemented on top of Transformer based pre-trained Language Models. Rather than a system itself, it could be considered as a methodology followed in order to scale fine-tuned LMs to its maximum performance. The first and main stage of the workflow implemented is searching and selecting a set of pretrained language models that better fit the requirements of the task. Once the base LMs have been selected, the focus turns into the corpus provided for the task, whose texts are the same for both subtasks faced. To enhance the number of instances that will be used in the fine-tuning phase of the transformer-based models, a common data augmentation strategy in NLP has been applied. Afterwards, the objective is centred into searching the best performing hyperparameters for each model and for each subtask. As well as the usual hyperparameters considered in deep learning networks, the cut-off threshold has been analysed. In the final stage, a basic ensemble method is considered combining the predictions of base estimators or models built in order to improve generalization capabilities and robustness over a single estimator.

3.1. Pretrained Language Models

Currently, Transformers are one of the leading deep learning technologies on the academy and the industry, and particularly on the NLP field. The core of Transformers are their attention mechanisms [2], a powerful computational block that works efficiently with sequential data, as text, trying to infer which parts of the sequence are more relevant within the context. Research and development on this type of architectures applied to a variety of NLP problems has increased at great speed in the last few years, given rise to platforms such as HuggingFace², which is utilized for the implementation of our system. HuggingFace is not just a framework of NLP models, but also a huge community of researchers and Artificial Intelligence enthusiasts.

A limited number of pretrained LMs have been tested as suitable candidates for both subtasks: the binary text classifier and the multilabel one. The main important characteristic is that the underlying language model supports the Spanish language. The following list collects the three models finally selected and its associated identifiers in the HuggingFace hub.

- BETO-Bert : dccuchile/bert-base-spanish-wwm-uncased
- Twitter-XLM-Roberta : cardiffnlp/twitter-xlm-roberta-base

¹<https://github.com/PlanTL-GOB-ES/lm-spanish>

²<https://huggingface.co>

- BSC-Roberta : BSC-TeMU/roberta-base-bne

Two specific-Spanish models are considered: BETO (with a size of BERT-base) [20] and BSC’s Roberta [21], both trained on large Spanish corpora such as the National Spanish Library (BNE, stand for “Biblioteca Nacional de España”). Additionally, a multilingual language model specialised in Twitter data [22] has been also elected for the given tasks.

3.2. Optimization techniques

In this section we briefly review the group of optimization techniques evaluated in our system approach.

3.2.1. Data augmentation

Since the size of the corpus has a considerable impact on the final model performance, one of our main concerns was to find or build a corpus with fairly enough quantity of text examples that make it possible for the transformer based models to grip the language subtleties that could be expressed on short tweets regarding the concepts of violence and aggressiveness. Based on the corpus released by the DA-VINCIS organization, which is composed of 3362 tweets manually annotated by at least 3 annotators, we have applied two types of transformations to increment the number of documents in the corpus.

The first is one of the most extended methods for textual data augmentation: the backtranslation. This approach consists in translating the original text or documents (in Spanish, in our problem) into some different languages and then translate back the result to the original one (Spanish). We relied on a multilingual encoder-decoder model trained for Many-to-Many multilingual translation developed by Facebook and released on the HuggingFace community: M2M100_1.2B³, selecting an intermediate translation language randomly from all supported languages.

The second transformation is based on more traditional text preprocessing techniques that are described in more detail in the section 3.3 below.

3.2.2. Hyperparameter tuning

Hyperparameter tuning consists of running multiple trainings of the model, varying its configuration parameters slightly in order to find the optimal combination that minimizes prediction error and maximizes objective metrics results automatically. The set and ranges of hyperparameters considered for tuning our system is depicted in the Table 1. Objective metrics contemplated are F1 score, precision and recall. The same sets of parameters and metrics have been used in both subtasks.

3.2.3. Cut-off calibration

The cut-off or probability threshold determines the value from which an observation is classified as positive (binary problems) or corresponds to a determined category (multi-class problems). It

³https://huggingface.co/facebook/m2m100_1.2B

Table 1
Hyperparameter exploration

Hyperparameter	Range
training epochs	5 to 10
batch size	4 to 16
learning rate	$1e^{-5}$ to $1e^{-3}$
weight decay	0.01 to 0.1
warmup ratio	0.05 to 0.2
adam epsilon	$1e^{-8}$ to $1e^{-6}$
gradient accumulation steps	1 to 2
max. number of trials	25

can be considered as one extra hyperparameter of the model, as it has a significant impact into the classification results. We have examined the performance of each model on each subtask for a range of probabilities between 0.2 and 0.9, in order to find the optimal cut-off value that improves the F1 scores.

3.2.4. Voting

Ensembling methods combine predictions from more than one machine learning algorithms in order to obtain better predictive performance than those algorithms separately. There are numerous approaches to tackle the optimal combination problem which differs in complexity and type of input features used to build the final model. In our system we have implemented a standard majority voting (majority out of three votes), the simplest efficient consensus algorithm usually followed in manual annotation procedures [23]. It considers the predictions made from every finetuned and optimized transformer-based models and decides the final prediction as the one that has appeared most often, giving equal importance to all models.

3.3. Data preprocessing

In addition, given the nature of the documents analysed, we have explored the possibility of applying some traditional yet simple text preprocessing techniques which may complete the context supplied to the models. Although Transformer-based LMs do not usually require prior data preprocessing, given the powerful tokenizers that they are built on, the documents collected from social media networks (in this particular case Twitter), present some elements habitually used by their users that might carry relevant information for the classification tasks [24]. We have analysed some of them in detail. **Links**: most of the tweets provided include shortened URLs that can be restored and consulted, as a first approximation we have crawled all the links and identified whether they belong to Twitter, external websites as well as if they contain media resources, replacing those URLs with literals providing that kind of information (i.e. “Enlace de publicación con foto” -post link with image-); **Hashtags**: these user-generated tagging expressions enable content cross-referencing on social platforms and also could convey intentions or key concepts of the text, we have split hashtags onto its individual tokens and normalize them; **Emojis**: small icons used to express an idea or emotion. Many of the instances

of the corpus contain at least one emoji that is closely related with the text and emphasizes its meaning. Emojis have different representations in different media platforms, but all have a corresponding unified definition⁴ in English. By integrating Demoji Python package⁵ into the system and a specific English-Spanish translator from the HuggingFace hub (*Helsinki-NLP/opus-mt-en-es*), we have replaced automatically all those icons with their standard description in Spanish.

4. Experimental setting and results

4.1. Experimental setup

With the goal of providing an experimental setup for reproducing the results exposed in this section, we include a description of the datasets and model configuration parameters used in the course of our trials.

Augmented training dataset: The original training and validation corpus released for DA-VINCIS task consisted of 3362 tweets manually annotated for both subtasks: 0,1 for violent event identification and one-hot-encoded for category recognition under the five labels described in the introductory section 1. Moreover, a trial set was initially released with a set of 50 tweets annotated under the same conditions. Train and trial sets were preprocessed as described above in the system description section (3.3) and a subset of that documents was augmented by means of backtranslation, building finally a training-validation corpus of 6874 examples. This dataset was split into training and validation subsets during training and optimization phases in a 90-10 ratio, having 688 text documents as validation reference used to compare results during the development.

Hyperparameter optimization: Hyperparameter optimization was executed over Nvidia GPUs (Tesla V100) on the maximum number of trials configured. Optimal results obtained for each subtask and model are shown in tables 2 and 3.

Cut-off calibration: Probability threshold analysis has been applied over the validation split for each of the optimized models with the aim of improving F1 scores. As illustrated in figures 1 for the binary classification and 2 for the multilabel problem, those scores do not fluctuate extremely, but adjusting the cut-off value adequately could have a positive impact on the final model performance.

In the case of multilabel classification, whether after applying the corresponding threshold, any instance remained uncategorized under any of the five classes, the text is assigned to the class with the highest output probability.

⁴<https://unicode.org/emoji/charts/full-emoji-list.html>

⁵<https://github.com/bsolomon1124/demoji>

Table 2

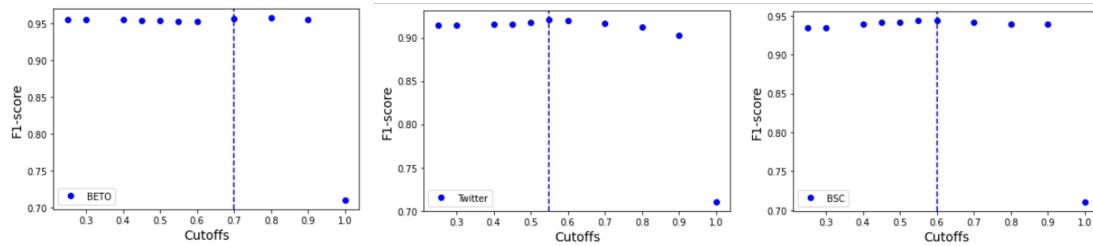
Best hyperparameters found for subtask 1

	BETO-Bert	Twitter-XLM-Roberta	BSC-Roberta
training epochs	7	7	7
batch size	15	12	11
learning rate	$8.9e^{-5}$	$2.77e^{-5}$	$1.15e^{-5}$
weight decay	0.0885	0.0134	0.0749
warmup ratio	0.065	0.079	0.183
adam epsilon	$7.25e^{-7}$	$7.29e^{-7}$	$2.95e^{-7}$
gradient accumulation steps	2	1	1
cutt-off	0.7	0.55	0.6

Table 3

Best hyperparameters found for subtask 2

	BETO-Bert	Twitter-XLM-Roberta	BSC-Roberta
training epochs	6	9	6
batch size	8	8	8
learning rate	$2.97e^{-5}$	$6.38e^{-5}$	$4.21e^{-5}$
weight decay	0.0584	0.0798	0.0929
warmup ratio	0.1471	0.0830	0.1105
adam epsilon	$7.31e^{-7}$	$1.38e^{-7}$	$5.57e^{-7}$
gradient accumulation steps	1	2	2
cutt-off	0.6	0.55	0.8

**Figure 1:** Optimal cut-offs for subtask 1

4.2. Results

In this section we expose the validation and test results obtained from the previously described experimental setup. Firstly, we have built the whole augmented training dataset. Later, it has been conveniently split and the same validation subset has been used on all runs with the aim of obtaining fair comparisons. Afterwards, the hyperparameter and cut-off optimization have been executed on the models for both subtasks independently. The metrics for the best runs found for each model, including the voting ensemble, and for each subtask during the development phase with our validation dataset are shown in the following tables 4 and 5. Additionally,

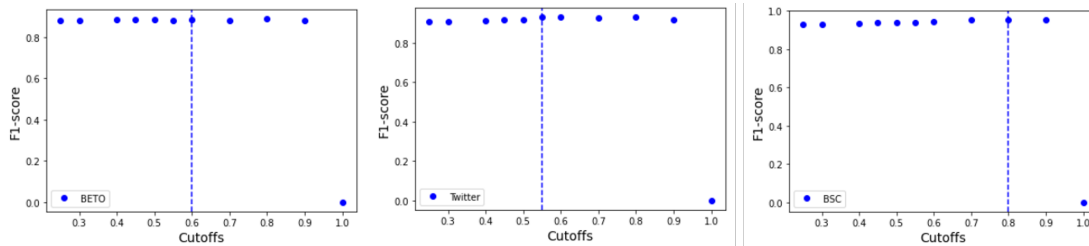


Figure 2: Optimal cut-offs for subtask 2

validation metrics obtained in an initial optimization phase, where just the original dataset (without augmentation) was used during the hyperparameter optimization, have been included in these tables.

We have also checked the effect of preprocessing input texts over the best individual model for each subtask, that is, the optimized BETO in binary classification and the optimized BSC-Roberta in the multilabel one. As the improvements are barely significant and unequal in both cases, we have decided not to include that preprocessing step for the model ensemble.

Table 4

Validation results for subtask 1

Model	F1	Precision	Recall
Voting	0.95017	0.94935	0.95126
Optimized BETO	0.94983	0.95251	0.94797
Optimized BSC Roberta	0.93851	0.93750	0.94010
Optimized Twitter XLM	0.91369	0.91268	0.91558
Optimized Beto + Preprocessing	0.95132	0.95377	0.94959
<i>Not augmented dataset</i>			
Optimized BETO	0.76774	0.78289	0.75317
Optimized BSC Roberta	0.75949	0.75949	0.75949
Optimized Twitter XLM	0.76191	0.68341	0.86076

An in deep analysis of validation predictions on the binary classification subtask shows that some tweets result difficult to classify for our system, as all the optimized models return the same predictions, and in some other cases there is one of them that is able to identify the aggressiveness yet the majority voting strategy leads to an incorrect result as depicted in figure 3.

As for multilabel text classification, misclassification issues are more variate: usually, at least one of the models tags the text correctly, but the others tend to fall into the “non-violent-incident” class (the majority class in this subtask annotations); there are also some difficulties in identifying more than one category when needed.

	text	beto	twit-xlm	bsc-rob	vote	true_labels
Revisa los siguientes consejos de seguridad y protégete de la nueva modalidad de robo de vehículos. 🚗 Ante cualquier sospecha, comunícate al 911. https://t.co/pTWHaamTcv		1	0	0	0	1
Accidente en C/ 14 #ZoomRadio919 #TraficoCR . Tráfico avanzando20m más lento de lo habitual. https://t.co/ft5uWb2OdE https://t.co/qD8XJF8Ww7		0	1	1	1	0
El presunto autor intelectual del asesinato de Yanelis Arias con 'ácido del diablo' https://t.co/8qDxTINK7W https://t.co/E88FyDQkxJ		0	0	0	0	1
Accidente en Ruta 27 / Próspero Fernández #ZoomRadio919 #TraficoCR . Tráfico avanzando32m más lento de lo habitual. https://t.co/AHKvvATPVm https://t.co/OFdlFHfz2F		0	0	1	0	1
Detención en flagrancia a dos sujetos por robo a usuarios del transporte público en Ecatepec https://t.co/ZYKj9GewNq https://t.co/1tMfrKcrUT		1	1	1	1	0

Figure 3: Validation errors subtask 1

Table 5

Validation results for subtask 2

Model	F1	Precision	Recall
Voting	0.94415	0.96086	0.93151
Optimized BSC Roberta	0.94265	0.94946	0.93865
Optimized Twitter XLM	0.94119	0.95187	0.93296
Optimized BETO	0.87182	0.88412	0.87167
Optimized BSC-Roberta + Preprocessing	0.94227	0.95647	0.93086
<i>Not augmented dataset</i>			
Optimized BETO	0.484943	0.664943	0.437715
Optimized BSC Roberta	0.55919	0.622715	0.528236
Optimized Twitter XLM	0.484442	0.547108	0.448541

Best performing models from previous experiments have been up-trained on the whole augmented dataset using the most suitable hyperparameters, then they have been used to make predictions on the test set which is composed of 1344 unseen tweets. Official results obtained either on binary as well as on multilabel classification are depicted on tables 6 and 7.

Table 6

Test results for subtask 1

Model	F1	Precision	Recall
Voting	0.76512	0.7792	0.75154
Optimized BETO	0.75735	0.7616	0.75316
Optimized BSC Roberta	0.75285	0.7920	0.71739
Optimized Beto + Preprocessing	0.75198	0.7568	0.74723
Optimized Twitter XLM	0.74863	0.7648	0.73312

text	text_labels	majority_txt	beto_txt	twitter_txt	bscrob_txt
Si al quemar #ArtificiosPirotecnicos se genera algún #Accidente, aléjate del perímetro y repórtalo inmediatamente a las autoridades correspondientes. https://t.co/gO0yruZzmx	non-violent-incident	accident	accident	non-violent-incident	accident
La sanción de la #FIA a #MaxVerstappen tras su accidente con #Hamilton: https://t.co/DiPRDGEFUG https://t.co/Ja8MlOW9yZ	accident	non-violent-incident	non-violent-incident	accident	non-violent-incident
Socorristas de @CruzRojaSal atendieron hoy a 2 personas que resultaron golpeadas en un accidente de tránsito, sobre el km 35 de la carretera troncal del norte en la jurisdicción de Aguilares. @NeryReyes https://t.co/hrcvKU4yGJ	robbery	accident	accident	accident	accident
Detención en flagrancia a dos sujetos por robo a usuarios del transporte público en Ecatepec https://t.co/ZYKj9GewNq https://t.co/11MfrKcrUT	non-violent-incident	robbery	robbery	robbery	robbery
Accidente en Bogotá-Mosquera / RN50-08A >Este #traficobogota. Tráfico hacia adelante1h 7m más lento que normalmente. https://t.co/Bv3qqTFD8B https://t.co/1dFkioxjK	non-violent-incident	accident,non-violent-incident	accident,non-violent-incident	non-violent-incident	accident,non-violent-incident
#TransitoyVialidad · Esta mañana se atiende accidente de tránsito en lateral de la carretera a Rioverde a unos metros de la Avenida de los Pinos. Masculino que posiblemente fue arrollado por un vehículo que se dio a la fuga. El atropellado se encuentra sin signos vitales. https://t.co/7Mm4rgxBG	accident,homicide	accident	accident	accident	accident,homicide

Figure 4: Validation errors subtask 2

Table 7

Test results for subtask 2

Model	F1	Precision	Recall
Optimized Twitter XLM	0.50459	0.50925	0.50373
Voting	0.50192	0.48281	0.53429
Optimized BSC Roberta	0.49620	0.50260	0.49076
Optimized BSC Roberta + Preprocessing	0.48417	0.48861	0.48102
Optimized BETO	0.46060	0.45213	0.47450

5. Conclusions and future work

The task of identifying aggressive and violent incidents on social media posts in Spanish language is quite challenging, not only due to the difficulty of working with the particular expressions and heterogeneous language used in Twitter, but also by the nature of the task itself. Sometimes identifying in a text a violent or aggressive incident could be rather subjective and in some other cases it could refer to some kind of sports event or films or being merely informative within the given context. As some specific conclusions derived from our design and experiments we can conclude that in the first subtask, violent event identification, it seems that the penalizations in our final scores come from the lower precision obtained (against top 2 participants), our recall is moderately better, so that a fair enough number of harmful events would be identified; while in the second subtask, violent category recognition, which is in general a more difficult task considering its multilabel nature and results on test data. This subtask could have been tackled as a 4-class classification on positive instances from the previous subtask in order to reduce the output dimension and yet its complexity. Using an ensemble strategy, as simple as voting in our case, generally improves individual results, thus exploring more robust approaches would be convenient. In contrast, preprocessing does not really improve performance (slightly in validation, but not in test), so that more complex preprocessing and more specific experiments

should be carried out in order to extract meaningful insights.

As future work lines we expect to explore further the different components of our approach by exploring larger, multilingual and social media specific language models, as well as particular Twitter preprocessing pipelines. Also, by doing an in deep assessment of data augmentation techniques considered, to verify that their application has had a real positive impact and has not led to model overfitting. Improvement of the preprocessing techniques applied should be contemplated by means of, for example, crawling information from links and applying multimodal models in case of having additional media attached to the post. Hashtag transformation could support the augmentation process by including synonyms or related tweets. A more detailed study on the use of emoticons would be interesting as well. Finally, exploit more sophisticated ensemble algorithms with different input features and model balancing in order to reinforce those that returned better metrics and mitigate the errors among them.

Acknowledgments

This work has been partially funded by the Department of Big Data and Cognitive Systems at the Technological Institute of Aragon, by IODIDE group of the Government of Aragon, grant number T1720R and by the European Regional Development Fund (ERDF).

References

- [1] L. J. Arellano, H. J. Escalante, L. Villaseñor-Pineda, M. Montes, G. F. Sanchez-Vega, Overview of da-vincis at iberlef 2022: Detection of aggressive and violent incidents from social media in spanish., *Procesamiento del Lenguaje Natural* 69 (2022).
- [2] 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).
- [3] L. Floridi, M. Chiriatti, Gpt-3: Its nature, scope, limits, and consequences, *Minds and Machines* 30 (2020) 681–694.
- [4] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, et al., A generalist agent, *arXiv preprint arXiv:2205.06175* (2022).
- [5] M. Baigang, F. Yi, A review: development of named entity recognition (ner) technology for aeronautical information intelligence, *Artificial Intelligence Review* (2022) 1–28.
- [6] M. del Carmen Rodríguez-Hernández, R. del-Hoyo-Alonso, S. Ilarri, R. M. Montañés-Salas, S. Sabroso-Lasa, An experimental evaluation of content-based recommendation systems: Can linked data and BERT help?, in: *17th IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2020, Antalya, Turkey, November 2-5, 2020*, IEEE, 2020, pp. 1–8. URL: <https://doi.org/10.1109/AICCSA50499.2020.9316466>. doi:10.1109/AICCSA50499.2020.9316466.
- [7] R. M. Montañés-Salas, R. del-Hoyo-Alonso, R. Aznar-Gimeno, From recurrency to attention in opinion analysis: Comparing RNN vs transformer models, in: M. Á. G. Cumbreñas, J. Gonzalo, E. M. Cámara, R. Martínez-Unanue, P. Rosso, J. Carrillo-de-

- Albornoz, S. Montalvo, L. Chiruzzo, S. Collovini, Y. Gutiérrez, S. M. J. Zafra, M. Krallinger, M. Montes-y-Gómez, R. Ortega-Bueno, A. Rosá (Eds.), Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019, volume 2421 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 589–597. URL: http://ceur-ws.org/Vol-2421/TASS_paper_4.pdf.
- [8] A. Singh, G. S. Josan, Paraphrase generation: A review from rnn to transformer based approaches., *International Journal of Next-Generation Computing* (2022).
- [9] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, arXiv preprint arXiv:2001.08361 (2020).
- [10] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al., Training compute-optimal large language models, arXiv preprint arXiv:2203.15556 (2022).
- [11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [12] 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).
- [13] Z. Li, E. Wallace, S. Shen, K. Lin, K. Keutzer, D. Klein, J. Gonzalez, Train big, then compress: Rethinking model size for efficient training and inference of transformers, in: *International Conference on Machine Learning*, PMLR, 2020, pp. 5958–5968.
- [14] N. Alnazzawi, Using twitter to detect hate crimes and their motivations: The hatemotiv corpus, *Data* 7 (2022) 69.
- [15] C. Shorten, T. M. Khoshgoftaar, B. Furht, Text data augmentation for deep learning, *Journal of big Data* 8 (2021) 1–34.
- [16] B. Li, Y. Hou, W. Che, Data augmentation approaches in natural language processing: A survey, *AI Open* (2022).
- [17] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, *Procesamiento del Lenguaje Natural* 68 (2022) 39–60. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405>.
- [18] J. Briskilal, C. Subalalitha, An ensemble model for classifying idioms and literal texts using bert and roberta, *Information Processing & Management* 59 (2022) 102756.
- [19] Z. Miftahutdinov, I. Alimova, E. Tutubalina, Kfu nlp team at smm4h 2019 tasks: Want to extract adverse drugs reactions from tweets? bert to the rescue, in: *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, 2019, pp. 52–57.
- [20] J. Canete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, *Pml4dc at iclr 2020* (2020) 2020.
- [21] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, M. Villegas, Spanish language models, 2021. arXiv:2107.07253.
- [22] F. Barbieri, L. Espinosa-Anke, J. Camacho-Collados, XLM-T: Multilingual Language Models

- in Twitter for Sentiment Analysis and Beyond, in: Proceedings of LREC, 2022.
- [23] J. Zhang, V. S. Sheng, Q. Li, J. Wu, X. Wu, Consensus algorithms for biased labeling in crowdsourcing, *Information Sciences* 382 (2017) 254–273.
- [24] A. Glazkova, M. Glazkov, T. Trifonov, g2tmn at constraint@ aaii2021: exploiting ct-bert and ensembling learning for covid-19 fake news detection, in: International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation, Springer, 2021, pp. 116–127.