Team GPLSI at AuTexTification Shared Task: Determining the Authorship of a Text

Iván Martínez-Murillo, Robiert Sepúlveda-Torres, Estela Saquete, Elena LLoret and Manuel Palomar

GPLSI research group, Dept. of Software and Computing Systems, University of Alicante, Ctra. San Vicente s/n, 03690, San Vicente del Raspeig, Alicante, España,

Abstract

AuTexTification is a shared task within the IberLEF workshop which aims to determine whether a text has been generated by an Artificial Intelligence (AI) or a human. The objective of this paper is to report the participation and results of the GPLSI team from the University of Alicante (Spain) in *subtask 1: Human or Generated* of the AuTexTification challenge for English and Spanish languages. We propose and experiment with different approaches based on Transfer Learning; Ensemble Learning; fine-tuning existing language models, such as RoBERTa or RemBERT; or relying on linguistic features. Our best models for both languages were trained through Transfer Learning techniques, obtaining the 6th and 8th position in the English and Spanish versions of this subtask, respectively. Results obtained in the Spanish-version were close to the top-performing team.

Keywords

Human Language Technologies, Transformers, Fine-tunning, Multilinguality, Ensemble classification, Transfer Learning

1. Introduction

In recent years, Natural Language Processing (NLP) has advanced exponentially, partly due to the development of Transformers [1] and Large Language Models (LLMs) [2]. Specifically, the task of Natural Language Generation (NLG) has benefited from this development. Thanks to this, some generative models such as GPT4 [3], PALM [4], or BLOOM [5] have been deployed with the ability to produce texts that, in some cases, can be indistinguishable from human-generated ones. Nevertheless, this rapid development also introduced some risks. On the one hand, even though automatically generated texts can be semantically correct and written in a human-like style, the meaning of the message, as well as the message itself, may be inaccurate or not true, which leads to hallucinations [6]. Furthermore, bias can be introduced in some cases during the training of these models, which could induce to produce unethical content [7]. On the other hand, it could be difficult to differentiate an original work from one generated by a machine in the academic context. In either of these cases, bad and unethical uses of Artificial Intelligence (AI) and its related technology could be promoted, for instance, to potentially generate misinformation.

IberLEF 2023, September 2023, Jaén, Spain

ivan.martinezmurillo@ua.es (I. Martínez-Murillo); rsepulveda@dlsi.ua.es (R. Sepúlveda-Torres); stela@dlsi.ua.es (E. Saquete); elloret@dlsi.ua.es (E. LLoret); mpalomar@dlsi.ua.es (M. Palomar)

Given this context and taking these problems into account, in recent years, the task of automatically detecting generated text has gained importance. Some proposals have been made, such as the AI Text Classifier [8] or GPTZero [9]. However, these approaches are not completely reliable yet. In order to fill this gap, *AuTexTification shared task* [10] has been proposed within the 2023 IberLEF workshop [11] for advancing the state of the art regarding the automatic detection of generated text in English and Spanish through two different subtasks.

- ***Subtask 1: Human or Generated*** is a binary text classification task in which it has to be determined whether a given text has been generated by an AI or not.
- ***Subtask 2: Model Attribution*** is a multiclass text classification task that consists in determining which AI model has generated a given text.

The objective of this paper is to report our participation (Team GPLSI) in the subtask 1 of AuTexTification - IberLEF 2023 shared task for English and Spanish languages. We will present the different models we have developed for automatically detecting AI-generated texts. We make use of Transfer Learning models and Ensemble Learning techniques, fine-tuning existing language models and linguistic feature extraction to address subtask 1 for both languages.

2. Background

AI Text detection is a task that has been studied for a while. Nevertheless, it has not been until recently, when it has really gained great relevance due to the explosion of generative AI. Particularly, in the context of NLG, despite its advantages, tools such as ChatGPT or Google Bard can pose some threats to society if they are not used critically or ethically. These tools generate texts very quickly, hardly indistinguishable from human-written texts, on top of a wide variety of styles and languages. Considering this, in some sectors, such as academia [12] or medicine [13], the need to detect if a text has been generated by an AI has arisen.

In light of this, there have been efforts in the research community to address this issue with the aim of advancing the state of the art. Some shared tasks related to automatic generated text detection involving different scopes have been proposed (e.g., DAGPap22 [14] or RuATD-22 [15]). The best model in DAGPap22 shared task was an ensemble of several fine-tuned models based on BERT [16], while in RuATD-22, the best models were those fine-tuned and based on BERT with extra features such as tonality, reading ease, or lexical richness [17].

In this context, AuTexTification shared task emerges as part of the IberLEF 2023 workshop. The aim of IberLEF is to promote research in text processing, understanding, and generation tasks in at least one of the Iberian languages. This 5^{th} edition includes the AuTexTification shared task, which as explained in Section 1, consists in automatically spotting if a text has been generated by an AI or by a human (subtask 1) and detecting which AI model has generated a given text (subtask 2).

3. AuTexTification Subtask 1 Overview: Human or Generated text

In this section, we provide a detailed description of subtask 1 of the AuTexTification shared task. Given a text as input, the aim of this subtask is to classify whether it has been generated by a human or a machine. This subtask can be addressed for two languages: English and/or Spanish, having the restriction that participants are only allowed to use the data given by the organisers to train the models.

In the next section, an explanation of the train and test datasets is provided.

3.1. Datasets

In the first stage of the shared-task, only training data was provided, so the participants could train and propose their approaches. This training data was provided in two distinct files, one for each language. Unlabelled test data was also provided, which was then labelled at a second stage, when the shared-task ended.

The remaining of this section describes the training and test datasets in both languages.

The first characteristic that is worth mentioning about the datasets is that, in some cases, they are not complete. Therefore, this task involves an additional challenge because it is not possible to make a distinction between AI-generated texts and human-written texts attending to grammatical criteria. An example of this issue can be seen in Table 1.

Table 1
Texts examples from training dataset.

Label	Spanish	English
Human	@user @user @user @user Si es así el spoiler, me va a emocionar ver esa localización en concreto	Let your friend know that you have noticed. This can encourage your friend to
Generated	@user por mensa Yo supuse que eras tú ¡Hola! ¿Cómo estás? ¿Estoy contento de que hayas adivinado qui	Laying down by the West in, One World Media was the first building to go up, followed by the Westin, a

Another fact to underline is the distribution between human and AI-generated texts. Table 2 shows the division of the dataset. It can be highlighted that the Spanish test dataset is not as balanced as the training dataset. In contrast, the English datasets are more balanced.

Table 2
Dataset texts distribution.

	Spanish			English		
Dataset	Human	Generated	Total	Human	Generated	Total
Train	15,787	16,275	32,062	17,046	16,799	33,845
Test	8,920	11,209	20,129	10,642	11,190	21,832

Moreover, analysing the length of these texts (number of characters) shown in Table 3 we

can see that AI-generated texts tend to be longer than human-generated texts for both datasets in English and Spanish. As well, the test dataset contains longer texts than the training one, regardless of the type (i.e., human or generated).

Table 3
Average text lenght (number of characters) for AuTexTification Subtask 1 datasets.

	Spanish		English	
Dataset	Human	Generated	Human	Generated
Training	297.12	316.67	297.13	313.49
Test	357.36	375.00	326.36	369.05

Finally, another aspect to highlight is the domain of texts in both datasets in both languages. While texts in the training dataset include several scopes such as legal documents, how-to articles, and social media, texts in the test dataset change the domain to news and reviews.

4. Team GPLSI Strategy

In this section the methodology followed to address subtask 1 is outlined.

4.1. Preliminary Statistical analysis

Based on existing literature [18], there are some patterns in a machine-generated text that could help to distinguish them from a human-written one. Particularly, text generated by a machine tends not to express sentiments and use more uncommon words. Relying on these findings, we decided to first conduct an analysis of the presence of some readability and sentiment features in the training dataset:

- To measure the **readability and understandability**, we used the textstat library (https://pypi.org/project/textstat/). The hypothesis is that AI-generated texts should be more difficult to read and understand. We employed four different functions to calculate statistics from the training dataset. Obtained results can be seen in Table 4.

Table 4
Mean of the scores for Readability and Understandability metrics (N/A indicates not applied for that language).

	Flesch Reading		**Szigriszt-Pazos**		**Fernandez-Huerta**		**McAlpine Eflaw**	
	The higher the better						The higher the worse	
Label	Spanish	English	Spanish	English	Spanish	English	Spanish	English
Human	78.18	77.01	74.25	107.45	78.08	N/A	N/A	22.04
Generated	74.73	74.04	70.79	104.58	74.64	N/A	N/A	25.76

1. The readability of a text can be assessed using the **Flesch Reading** Ease formula [19] (a higher score means that a text is easier to understand).

2. The understandability of a text can be rated by applying the **Szigriszt-Pazos** formula [20] (a higher the score is indicates that a texts is easier to understand).
3. **Fernandez-Huerta** [21] formula measures the readability of the text, but this metric is specifically used to evaluate texts in Spanish (a bigger value denotes that a text is easy to read).
4. To calculate the readability of an English text by a foreign learner, we used the **McAlpine EFLAW** Readability Score [22] (the lower the score of a text, the easier it is to read for a foreigner person.). This metric is specifically used for the English dataset.

Based on the results obtained in Table 4, we can extract the conclusion that human-written texts tend to be easier to read and understand than AI-generated ones. Based on this finding, we consider that using these metrics will help to increase results for this task.

- We calculated **sentiment** statistics of the training dataset using the NLTK library [23]. The hypothesis to use this is that texts generated by an AI tend to be more neutral than texts written by humans.
  Based on the findings of Table 5, we can see that in Spanish, most texts do not express sentiment, being tagged as neutral. Analysing the other sentiment categories, there are more texts expressing a positive sentiment in human-written texts than in AI-generated texts.
  In contrast, the results obtained for the English dataset were not the expected ones. As we can see in Table 5, most of the texts express a positive sentiment for both human and AI generated texts, so texts generated by an AI can also be trained to express sentiments. On the contrary, there is a greater number of human-generated texts expressing a negative sentiment than texts generated by an AI in the English dataset.

Table 5
Sentiment detection in the AuTexTification subtask 1 training dataset.

	Spanish			English		
Label	Positive	Negative	Neutral	Positive	Negative	Neutral
Human	3,178	5,870	6,739	9,570	4,621	2,855
Generated	2,328	6,510	7,437	11,065	3,458	2,276

4.2. Base Models

In our research, Transfer Learning techniques have been applied. Transfer Learning is a subfield of machine learning that applies knowledge learned by solving one task to approach a different task [24]. In this sense, we have used pre-trained models based on Transformers architecture that obtains relevant results in NLP tasks. These models are trained in a general task and can be fine-tuned on a more specific task. For this purpose, the model is set into a fine-tuning mode. Afterwards, training examples are used to adjust the weights of a language model and the neural network classifier, which makes the final prediction. Two types of pre-trained models have been used to address this task, specific language models and multilingual models. The employed models are described below:

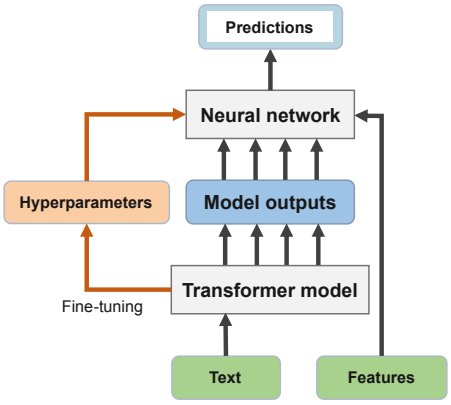
- **Specific language model:** We use RoBERTa and BETO models trained in English and Spanish. These are based on BERT, maintaining the same structure but optimising a few parameters, such as larger training data or larger batch size. These optimisations make RoBERTa performs better than BERT in the English language. Specifically, we use the following models to address the English and Spanish tasks:
  - English: RoBERTa-base-openai-detector was trained to detect GPT-generated text. Further information can be found in [25]. This model is published at https://huggingface.co/roberta-base-openai-detector.
    RoBERTa-large has been pretrained with a collection of five different datasets. These datasets have a total weight 160GB of text. More information, as well as the model itself, is publicly available at https://huggingface.co/roberta-large [26].
  - Spanish: RoBERTa-base-bne was trained with a total of 570GB of clean Spanish texts. Further information can be obtained in [27]. This model is accesible in https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne.
    BETO was instructed with texts in Spanish from Wikipedia and the OPUS project (https://opus.nlpl.eu/). More information is provided in [28]. This model is published at https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased.
- **Multilingual model:** We use XLM-RoBERTa and RemBERT models trained on a multilingual dataset. These models can be trained for specific tasks in different languages.
  XLM-RoBERTa is a multilingual model trained with 2.5TB of filtered CommonCrawl data over 100 languages [29]. This model can be found in https://huggingface.co/xlm-roberta-base.
  As well as XLM-RoBERTa, RemBERT is also a multilingual model which can be seen as a bigger version of mBERT [30]. This model is trained with 26B tokens of Wikipedia data over 110 languages and is publicly available in https://huggingface.co/google/rembert.

4.3. Base models with features

Some works include additional features to combine them with the output of the last layer of the transfer learning models [31, 32, 33]. This strategy could improve the prediction of models based on transfer learning. In this sense, we propose to include the features explained in section 4.1 because the statistical analysis shows differences between human and generated texts. We relied on the approach proposed in [33]. Figure 1 shows the internal architecture of this approach.

The input text was encoded by using a transformer model. The encoded vector (model output) is concatenated with the features in the first layer of the neural network. Afterwards, a dropout layer is applied to prevent overfitting, and finally, an output layer classifies between human and generated text. The output layer, in this case, is a dense layer with two output neurons using a softmax activation function and cross-entropy as a loss function. In this case, the neural network is a Multilayer Perceptron (MLP) with three layers. Our approach adopts the classification architecture proposed by [33], which combines Transformer model with external features.

Figure 1: Internal architecture of the classification system when uses features [33].



4.4. Ensemble Models

Based on the promising results obtained in other challenges [16], Ensemble models could improve the performance of some tasks. Particularly, we have used an Ensemble Stacking method. This approach consists in training some models to predict a given task and combining its predictions with training a meta-learner to output a final prediction. The meta-learner inputs the predictions of the stacked models as features, and the target label, and it learns how to best combine the input predictions to make a better output prediction by using a traditional machine learning algorithm [34].

5. Experimental Setup

In this section, the main experiments carried out to make the proposals for the workshop will be explained. During the training phase, only the training dataset was available. The training datasets for English and Spanish were, in turn, split into three subsets: one for training the models, other for validation, and the last one for testing the models. The validation subset was used to estimate model skills while performing hyperparameter tuning and to compare different approaches. The test subset was used to measure the performance and generalisation of the model with unseen data [35]. Each validation and test subset approximately represent a 12% of the training sets. Table 6 shows the distribution per class in each subset created.

Table 6
Training, validation, and test subsets (from the training dataset) distribution per class.

	English			Spanish		
Subsets	Human	Generated	Total	Human	Generated	Total
Training	13,563	13,241	26,804	12,414	12,978	25,392
Validation	1,811	1,845	3,656	1,730	1,733	3,463
Test	1,672	1,713	3,385	1,643	1,564	3,207

As can be seen in Table 6, all the subsets are balanced between human and generated classes.

To tackle the strategy defined in Section 4, four experiments have been proposed. These experiments are described below:

1. ***Monolingual dataset fine-tuning:*** We fine-tuned some state-of-the-art Spanish, English, and multilingual pre-trained models with an initial hyperparameter configuration. The initial hyperparameter configuration was a maximum sequence length of 400, a batch size of 8, a training rate of 1e-5, a manual seed of 1,509, and training epochs of 3. The models with better performance to predict the validation subsets and with less overfitting were used to perform a bayesian hyperparameter tuning. To automate hyperparameter tuning, we used the Weight & Biases library [36]. Table 7 shows the search configuration.

Table 7
Hyperparameter tuning.

Parameters	Values
Number train epochs	2, 3, or 4
Dropout	0.1, 0.2, or 0.3
Batch size	4, 8, or 16
Learning rate	1e-5, 1.5e-5, 2e-5, 2.5e-5, 3e-5, 3.5e-5, or 4e-5
Weight decay	0, 0.1, 0.2, or 0.3

2. ***Multilingual dataset fine-tuning:*** The second experiment concatenates English and Spanish training datasets to train multilingual models with Transformer architecture. This process can obtain more general models since more training data is available, and also multilingual models tend to be very stable since they find patterns beyond the features of each language.
3. ***Fine-tuning with features:*** Based on existing literature, adding features extracted from the training data, such as sentiment or readability, while training a model could improve its performance. Figure 1 shows the architecture proposed by Sepúlveda-Torres et al., that was used in this experiment. Six features were employed for each language: a general readability score (Flesch Reading Ease), a readability language specific score (McAlpine EFLAW and Fernandez-Huerta, English and Spanish, respectively), an understandability score, and three features that represent sentiments. Before concatenating the features to the outputs of the transformer model, a normalisation of the readability and understandability features were performed. We normalised these features since they were in a different range of values from the sentiment features.
4. ***Ensemble:*** The last experiment for both languages is an ensemble classification system, aiming to integrate the best obtained models and thus, improve prediction results. In this case, we used a Logistic Regression (LR) algorithm as meta-learner. The logit outputs of each model are stacked and used as predictors of LR algorithm. These logit outputs also have been normalised to perform the training and prediction.

6. Results and discussion

All of the experiments explained in Section 5 were implemented using *Simple Transformer* [37] and *PyTorch* [38] libraries. The data employed to train the models were only derived from the data provided by the challenge. The final hyperparameters configuration for each classifier and the code implemented can be found at https://github.com/rsepulveda911112/Autextification.

6.1. Training phase

This section shows the results of all the experiments carried out with the different Transformers models and the strategies addressed in section 4.

Table 8 and Table 9 show the results of the experiments for English and Spanish, respectively. Each row is annotated according to the experiment it represents. To evaluate the performance of each experiment, we have applied the official metrics of the AuTexTification shared task (Macro-F1 in percentage mode).

Table 8
IA vs. human prediction results for English on the validation and test subsets created by us.

No	Systems	Macro-F1 (%) Validation	Test
1	***Monolingual dataset fine-tuning*** (RoBERTa)	91.10	90.44
2	***Monolingual dataset fine-tuning*** (RoBERTa-base-openai-detector)	93.41	92.94
3	***Monolingual dataset fine-tuning*** (RemBERT **(Run_1)**)	**95.54**	**94.79**
4	***Multilingual dataset fine-tuning*** (XLM-RoBERTa-large concatenated datasets)	90.61	92.03
5	**Multilingual dataset fine-tuning** (RemBERT concatenated datasets **(Run_2)**)	**95.30**	**95.39**
6	***Fine-tuning with features*** (RoBERTa)	90.70	90.45
7	***Ensemble*** **(Run_3)**	-	**96.01**

Table 9
IA vs. human prediction results for Spanish on the validation and test subsets created by us.

No	Systems	Macro-F1 (%) Validation	Test
1	***Monolingual dataset fine-tuning*** (RoBERTa-base-bne)	95.09	95.13
2	***Monolingual dataset fine-tuning*** (BETO)	93.43	92.49
3	***Monolingual dataset fine-tuning*** (RemBERT)	94.97	95.03
4	***Multilingual dataset fine-tuning*** (XLM-RoBERTa-large concatenated datasets)	90.61	91.87
5	**Multilingual dataset fine-tuning** (RemBERT concatenated datasets **(Run_2)**)	**95.30**	**95.72**
6	***Fine-tuning with features*** (RoBERTa-base-bne with features **(Run_1)**)	**94.71**	**95.59**
7	***Ensemble*** **(Run_3)**	-	**96.66**

The first experiment in Tables 8 and 9 (rows 1, 2, and 3) obtains competitive results for the RemBERT model in English and for RoBERTa-base-bne and RemBERT models in Spanish. The results of this first experiment for these models are obtained after performing a hyperparameter search.

The second experiment (rows 4 and 5) achieves high results when using the RemBERT language model. The other model gets very discrete results.

In the case of the third experiment, results were not as good as expected. We obtained poor results when adding features to an English model, and adding features to a Spanish model performed similarly to without them.

Finally, the best results in both languages were obtained after using the ensemble model, in the same way as reported in published research works with similar approaches [16]. This approach significantly improved the results in the Spanish dataset. The English ensemble used **Run_1** and **Run_2** models. In the case of Spanish, the ensemble has been constructed with the following models:

- (Run_1) Roberta-base-bne with features.
- Roberta-base-bne with Hyperparameter tuning.
- (Run_2) RemBERT trained with datasets in both languages concatenated.
- RemBERT trained with the Spanish dataset.

Through the training phase, we noticed that models did not overfit when splitting the training dataset into three subsets, and results obtained when predicting both validation and test subsets were similar. On account of this, we considered that training our models with more data would increase our results. So, for making the final submission, we have only split the training data into two subsets (one for training, and the other for test). Moreover, the *Ensemble experiment* was performed after training the final models. Hence, we do not have validation results for the Ensemble experiment, as it could only be tested with unseen data.

These experiments carried out guided us to determine the three models to be finally submitted for the AuTexTification subtask-1, either for English or Spanish. The submitted models will be evaluated with the test set released by the subtask organisers.

6.2. Submission results: Subtask 1 - English

The three submitted models for subtask 1 in English can be seen in Table 10. As the best results while experimenting with the English dataset were making use of multilingual models, our **Run_1** and **Run_2** are a RemBERT multilingual model. The main difference is that in **Run_1** the model has been only trained with the English data, and in **Run_2** the model has been trained with both English and Spanish training data. Finally, **Run_3** is an ensemble of both models.

Table 10
Oficial subtask prediciton results for English.

Systems	F_1 Score (%) Generated	Human	Macro-F1 (%)
Run_1: ***Monolingual dataset fine-tuning*** (RemBERT)	**79.93**	**65.11**	**72.52**
Run_2: ***Multilingual dataset fine-tuning*** (RemBERT)	78.97	62.10	70.54
Run_3: ***Ensemble***	79.39	63.39	71.39

For the English subtask, our best model has been **Run_1**. It achieved the 6th position out of 76 proposals, with a score of 72.52 for the Macro-F1 metric. In contrast with the training phase,

Run_2 (the model trained with the multilingual dataset) did not perform as well as the model trained with just the English dataset reaching 11th place with a score of 70.54 in the Macro-F1 metric. Finally, despite the fact that the ensemble model obtained the best results in the training phase, **Run_3** did not obtain the best results with the test dataset. Nonetheless, it has achieved a 10th position in the ranking with similar results to our best approach (Macro-F1 of 71.39), so this means that it is also a good strategy.

Comparing these results with those obtained by other participants, the top-performing system in this subtask obtained 80.91, followed by the second best one, which obtained 74.16. Our best model was very close to the second one in the results.

6.3. Submission result: Subtask 1 - Spanish

For the Spanish subtask 1, we have also submitted three different models. In this case, **Run_1** has been a Spanish model, trained with the Spanish dataset concatenated with features. In the same way as for English, **Run_2** has been the multilingual RemBERT model trained with the datasets in both languages concatenated. Finally, the proposal for **Run_3** has been an ensemble built with four different models.

Table 11
Oficial subtask prediciton results for Spanish.

Systems	F_1 Score (%)		Macro-F1 (%)
	Generated	**Human**	
Run_1: ***Fine-tuning with features*** (RoBERTa-base-bne with features)	78.09	48.28	63.19
Run_2: ***Multilingual dataset fine-tuning*** (RemBERT)	**79.62**	**54.02**	**66.82**
Run_3: ***Ensemble***	78.55	49.23	63.90

Results obtained in the final submission can be seen in Table 11. Our best approach at predicting the Spanish task has been **Run_2**, achieving the 8th position in the ranking with a Macro-F1 of 66.82. The multilingual model trained with the dataset concatenated in both languages performed satisfactorily for this task. In contrast, extracted features seem not to be relevant to predict whether a text has been generated by an AI or a human with the given test dataset because **Run_1** model has not achieved the expected results (22nd position obtaining a Macro-F1 of 63.19). Finally, Ensemble did not perform as well as expected, considering that **Run_3** reached the 20th position with a Macro-F1 of 63.9, presumably because the ensemble was built with models in different languages (multilingual and Spanish only), and it may not be selecting the relevant features to predict just one language.

When comparing the results with those achieved by other participants in Spanish, the leading system in this particular subtask achieved a score of 70.77 in the Macro-F1 metric. In this case, our best model was very proximate to the first position model.

6.4. General discussion

The official results obtained after the submission of our approaches (Team GPLSI) show a noticeable difference compared to the results obtained in the experimentation. The reason

for that could be that our models have learned to predict well some specific domains (legal documents, how-to articles, and social media), but when changing the domain to predict unseen domains (news and reviews), they do not behave in the same way. However, obtained results indicate that all our models achieved a high score when predicting AI-generated text, being the assignment of predicting human written the one which not achieves as good results. This could happen because, as explained in Section 3.1, both human and AI generated texts are incomplete in some cases. In turn, this may provoke that models not to learn as well some grammatical patterns such as punctuation and consequently predict as an AI generated text one which is actually written by a human.

Another important issue that is worth discussing is that a multilingual model fine-tuned with a dataset in a specific language performs better than fine-tuning the same model with two different languages. Nonetheless, results of both models are close. Thereby, in some cases would be interesting to construct a model that could make multilingual predictions, being more efficient, than constructing a model that is only able to predict in one language.

Moreover, looking at the competition results we can see that, generally speaking, models predicting the Spanish dataset perform worse than the models trained to predict the English dataset. Such results could arguably be a consequence of the structural differences between both languages, as the English language tends to show more pre-fixed grammatical patterns with simpler structures, whereas in the case of Spanish, the grammatical constructions can be built with a wider variety of linguistic structures. Consequently, it could be more difficult for the models to cover every grammatical pattern concerning the Spanish language, which would cause the worse performance shown in the results.

7. Conclusions and Future Work

The task of automatically detecting texts generated by an AI is a growing task in the NLP field as a consequence of the rapid development of NLG. Current generative models can generate automatic texts that are hardly indistinguishable from human-written texts. This issue has caused some social concerns.

In this context, AuTexTification challenge was proposed as part of the IberLEF 2023 workshop. We (Team GPLSI) participated in subtask 1 for both languages, Spanish and English. The results achieved in the English subtask are considered good, ranking 6^{th} of 76 participants with a Macro-F1 of 72.52. In the Spanish subtask, we ranked 8^{th} of 52 participants with a Macro-F1 of 66.82. The interesting fact about this model is that it is a multilingual model fine-tuned with both datasets (English and Spanish) and achieved a high place in both tasks (8^{th} with a F1 of 66.82 in Spanish and 11^{th} with a F1 of 70.54 in English). This could confirm that fine-tuning a multilingual pre-trained model could be a good option to automatically detect AI-generated texts in several languages while being more efficient than building a model to predict just one language. Another aspect to highlight is that results predicting AI-generated texts are high, achieving almost a Macro-F1 of 80 in all of our models. However, these models do not predict human written texts as well as when it comes to AI-generated texts.

One future line of work is to analyse and extract other linguistic patterns in human texts to train the model and help it to better understand the way humans write, and consequently,

increase the results obtained at predicting it. In addition to this, we will experiment with other approaches, such as Active Learning or Generative Adversarial Networks to measure how well these approaches perform in comparison with the proposed models during this task. Furthermore, we will address these tasks with no restrictions on the data so we can train our models with more text domains with the objective to make a better generalisation of the knowledge and, consequently, predicting other domains better.

Acknowledgements

This research work is part of the R&D projects "CORTEX: Conscious Text Generation" (PID2021-123956OB-I00) and "TRIVIAL: Technological Resources for Intelligent VIral AnaLysis through NLP" (PID2021-122263OB-C22), both funded by MCIN/ AEI/10.13039/501100011033/ and by "ERDF A way of making Europe", and "CLEAR.TEXT:Enhancing the modernization public sector organizations by deploying Natural Language Processing to make their digital content CLEARER to those with cognitive disabilities" (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and "European Union NextGenerationEU/PRTR". Moreover, it has been also partially funded by the Generalitat Valenciana through the project "NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/21)", and by the European Commission ICT COST Action "Multi-task, Multilingual, Multi-modal Language Generation" (CA18231).

References

[1] 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).

[2] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023).

[3] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[4] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al., Palm 2 technical report, arXiv preprint arXiv:2305.10403 (2023).

[5] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).

[6] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38.

[7] J. Homolak, Opportunities and risks of chatgpt in medicine, science, and academic publishing: a modern promethean dilemma, Croatian Medical Journal 64 (2023) 1.

[8] OpenAI, 2023. URL: https://beta.openai.com/ai-text-classifier.

[9] E. Tian, Gptzero, Retrieved Jan 25 (2023) 2023.

[10] A. M. Sarvazyan, J. Á. González, M. Franco Salvador, F. Rangel, B. Chulvi, P. Rosso, Overview of AuTexTification at IberLEF 2023: Detection and Attribution of Machine-

Generated Text in Multiple Domains, in: Procesamiento del Lenguaje Natural, Jaén, Spain, 2023.
[11] 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).
[12] M. M. Rahman, Y. Watanobe, Chatgpt for education and research: Opportunities, threats, and strategies, Applied Sciences 13 (2023) 5783.
[13] L. De Angelis, F. Baglivo, G. Arzilli, G. P. Privitera, P. Ferragina, A. E. Tozzi, C. Rizzo, Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health, Frontiers in Public Health 11 (2023) 1567.
[14] Y. Kashnitsky, D. Herrmannova, A. de Waard, G. Tsatsaronis, C. Fennell, C. Labbé, Overview of the dagpap22 shared task on detecting automatically generated scientific papers, in: Third Workshop on Scholarly Document Processing, 2022.
[15] T. Shamardina, V. Mikhailov, D. Chernianskii, A. Fenogenova, M. Saidov, A. Valeeva, T. Shavrina, I. Smurov, E. Tutubalina, E. Artemova, Findings of the the ruatd shared task 2022 on artificial text detection in russian, arXiv preprint arXiv:2206.01583 (2022).
[16] A. Glazkova, M. Glazkov, Detecting generated scientific papers using an ensemble of transformer models, arXiv preprint arXiv:2209.08283 (2022).
[17] N. Maloyan, B. Nutfullin, E. Ilyushin, Dialog-22 ruatd generated text detection, arXiv preprint arXiv:2206.08029 (2022).
[18] S. Mitrović, D. Andreoletti, O. Ayoub, Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text, arXiv preprint arXiv:2301.13852 (2023).
[19] R. Flesch, A new readability yardstick., Journal of applied psychology 32 (1948) 221.
[20] F. Szigriszt Pazos, Sistemas predictivos de legilibilidad del mensaje escrito: fórmula de perspicuidad (1992).
[21] J. Fernández Huerta, Medidas sencillas de lecturabilidad, Consigna 214 (1959) 29–32.
[22] R. McAlpine, From plain english to global english, 2012.
[23] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, " O'Reilly Media, Inc.", 2009.
[24] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proceedings of the IEEE 109 (2020) 43–76.
[25] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, et al., Release strategies and the social impacts of language models, arXiv preprint arXiv:1908.09203 (2019).
[26] 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.
[27] 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.
[28] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[29] 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, arXiv preprint arXiv:1911.02116 (2019).

[30] H. W. Chung, T. Fevry, H. Tsai, M. Johnson, S. Ruder, Rethinking embedding coupling in pre-trained language models, arXiv preprint arXiv:2010.12821 (2020).

[31] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, X. Zhou, Semantics-aware BERT for language understanding, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 9628–9635. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6510.

[32] W. M. Lim, H. T. Madabushi, UoB at SemEval-2020 Task 12: Boosting BERT with Corpus Level Information (2020).

[33] R. Sepúlveda-Torres, M. Vicente, E. Saquete, E. Lloret, M. Palomar, Leveraging relevant summarized information and multi-layer classification to generalize the detection of misleading headlines, Data & Knowledge Engineering 145 (2023) 102176. URL: https://www.sciencedirect.com/science/article/pii/S0169023X23000368. doi:https://doi.org/10.1016/j.datak.2023.102176.

[34] Y. Khandelwal, Ensemble stacking for machine learning and deep learning, 2021. URL: https://www.analyticsvidhya.com/blog/2021/08/ensemble-stacking-for-machine-learning-and-deep-learning/.

[35] S. J. Russell, Artificial intelligence : a modern approach, fourth edition, Pearson series in artificial intelligence, global ed. ed., Pearson Education, Harlow, 2022.

[36] L. Biewald, Experiment tracking with weights and biases, 2020. URL: https://www.wandb.com/, software available from wandb.com.

[37] T. C. Rajapakse, Simple transformers, https://github.com/ThilinaRajapakse/simpletransformers, 2019.

[38] 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).