Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 421-434 recibido 04-07-2024 revisado 18-07-2024 aceptado 26-07-2024

Overview of IberAuTexTification at IberLEF 2024: Detection and Attribution of Machine-Generated Text on Languages of the Iberian Peninsula

Resumen de IberAuTexTification en IberLEF 2024: Detección y Atribución de Texto Generado por Máquina en Idiomas de la Península Ibérica

Areg Mikael Sarvazyan,[1] **José Ángel González,**[1] **Francisco Rangel,**[1]
Paolo Rosso,[2,3] **Marc Franco-Salvador**[1]
[1]Genaios, Valencia, Spain
[2]Universitat Politècnica de València, Valencia, Spain
[3]Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI)
{areg.sarvazyan, jose.gonzalez, francisco.rangel, marc.franco}@genaios.ai
prosso@dsic.upv.es

Abstract: This paper presents the overview of the IberAuTexTification shared task as part of the IberLEF 2024 Workshop in Iberian Languages Evaluation Forum, within the framework of the SEPLN 2024 conference. IberAuTexTification extends our previous AuTexTification shared task in three dimensions: (i) more domains, (ii) more languages from the Iberian Peninsula, and (iii) more prominent LLMs. This shared task frames a multilingual, multi-domain, and multi-model setting consisting of two subtasks. For Subtask 1, participants have to determine whether a text's author is a human or machine. For Subtask 2, participants have to attribute a machine-generated text to a large language model. Our IberAuTexTification dataset contains about 168,000 texts across six languages (English, Spanish, Portuguese, Catalan, Basque, and Galician) and seven domains (chat, news, literary, reviews, tweets, wikipedia, and how-to articles). A total of 21 teams participated in the task with 68 runs, 54 for Subtask 1 and 14 for Subtask 2. In this overview, we present the IberAuTexTification task, the submitted participating systems, and the results.
Keywords: IberAuTexTification, Languages of the Iberian Peninsula, Machine-Generated Text, Large Language Models.

Resumen: Este artículo presenta un resumen de la tarea IberAuTexTification como parte del workshop IberLEF 2023 en el Iberian Languages Evaluation Forum, dentro del marco de la conferencia SEPLN 2024. IberAuTexTification extiende nuestra tarea previa, AuTexTification, en tres dimensiones: (i) más dominios, (ii) más idiomas de la Península Ibérica y (iii) LLMs más destacados. Esta tarea propone un escenario multilingüe, multi-dominio y multi-modelo consistente en dos subtareas. En la Subtarea 1, los participantes deben determinar si el autor de un texto es un humano o una máquina. En la Subtarea 2, los participantes deben atribuir un texto generado al modelo de lenguaje que lo generó. Nuestro dataset de IberAuTexTification contiene alrededor de 168.000 textos en seis idiomas (Inglés, Español, Portugués, Catalán, Vasco y Gallego) y siete dominios (chat, noticias, literatura, reseñas, tweets, wikipedia y artículos instructivos). Un total de 21 equipos participaron en la tarea, enviando 68 resultados, 54 para la Subtarea 1 y 14 para la Subtarea 2. En este artículo, presentamos la tarea IberAuTexTification, los sistemas enviados por los participantes y sus resultados.
Palabras clave: IberAuTexTification, Idiomas de la Península Ibérica, Texto Generado por Máquina, Modelos de Lenguaje Masivos.

1 Introduction

There is a growing trend in the use of Machine-Generated Text (MGT) for all kinds of tasks. Easy access to Large Language Models (LLMs) such as GPT (Ouyang et al., 2022; Achiam et al., 2023), LLaMA (Touvron

ISSN 1135-5948 DOI 10.26342/2024-73-32

et al., 2023), Gemini (Team et al., 2023), and Mixtral (Jiang et al., 2024) is enabling non-technical people to generate human-like, high-quality, multi-style, and multi-domain content and solutions (Eloundou et al., 2024; Liu et al., 2023). However, this content could also threaten our society via intellectual property rights violations (Henderson et al., 2023), data leakage (Nasr et al., 2023), and malicious uses such as generating fake-news, polarised opinions, or smear campaigns (Kasneci et al., 2023). Therefore, there is a need for solutions to 1) **detect** MGT (*Is this text machine generated?*) and 2) **attribute** it to specific text generation models (*Which model generated this text?*).

The MGT detection has successfully been applied in isolation to specific domains, data sources, languages, or models (Bakhtin et al., 2019). However, there is room for improvement in real-world applications, where these variables are combined in large-scale scenarios (Eloundou et al., 2024). Recent studies include zero-shot approaches (Mitchell et al., 2023; Zellers et al., 2019) and supervised systems (Ippolito et al., 2020; Uchendu et al., 2020) for document-level detection and attribution. Other works have studied generalization across model families and scales (Sarvazyan et al., 2023b; Antoun, Sagot, and Seddah, 2024), and mixcase detection, with interleaved machine-generated and human texts (Zhang et al., 2024).

Encouraged by the popularity of LLMs and the need to foster research and knowledge sharing to detect their generations, several initiatives and evaluation campaigns have been organized. Some focused on the detection task and addressed the different languages separately, i.e., in multi-domain English texts (Molla et al., 2023), in English essays,Other tasks included additional tasks such as multi-domain model attribution in separate English and Spanish (Sarvazyan et al., 2023a), as well as multi-domain and multilingual model attribution and boundary detection tracks (Wang et al., 2024b). Additionally, recent shared tasks include efforts to understand and explain the behaviour of these detectors (Fivez et al., 2024). Finally, novel frameworks such as TextMachina (Sarvazyan, González, and Franco-Salvador, 2024b)[1] focus on the creation of high-quality MGT datasets, for any language and LLM, supporting all the mentioned tasks: detection, attribution, boundary detection, and mixcase.

In this paper, we present the **IberAuTexTification** (**Au**tomated **Tex**t iden**Tification** on languages of the **Iber**ian Peninsula) Shared Task[2] at IberLEF 2024 (Chiruzzo, Jiménez-Zafra, and Rangel, 2024). We extend our previous AuTexTification at IberLEF 2023 task (Sarvazyan et al., 2023a) in three dimensions: more text generation models, more domains, and more languages from the Iberian Peninsula. In addition, our detection and attribution tracks follow a multi-domain and multilingual setting, with special focus on cross-domain generalization to build robust detectors and attributors in real-world scenarios. This is, to the best of our knowledge, the first attempt to conduct the MGT detection and attribution tasks in languages from the Iberian Peninsula.

2 *Task Description*

The IberAuTexTification Shared Task includes two subtasks involving six languages from the Iberian Peninsula: Spanish 🇪🇸, Catalan, Basque, Galician, Portuguese 🇵🇹, and English 🇬🇧. It also encompasses seven different domains to cover a wide variety of writing styles: Chat 💬, How-to, News 📰, Literary 📚, Reviews, Tweets 🐦, and Wikipedia 🌐.

As opposed to the previous edition, here we consider a multilingual setting, where all six languages are mixed into a single dataset for each subtask. Moreover, to resemble real scenarios where domain generalization is crucial, in both subtasks we split the datasets by domain, i.e., the training sets include five domains, while the test sets comprise two unseen domains.

Subtask 1: MGT Detection. Consists in identifying whether a text's author is a human or LLM. Framed as a binary classification task with human 👩 and LLM 🤖 labels.

Subtask 2: MGT Attribution. Participants must attribute MGT to the LLM that generated it. Framed as a multi-class classification task with six labels, one for each LLM that generated MGT in our dataset.

[1] https://github.com/Genaios/TextMachina

[2] https://sites.google.com/view/iberautextification/

3 Dataset

The IberAuTexTification dataset includes texts written by humans and generated by LLMs in six languages and seven domains. To build such dataset, we gather human texts and use TextMachina (Sarvazyan, González, and Franco-Salvador, 2024b), a Python package to easily create datasets for MGT-related tasks. By employing TextMachina, we ensure that the final dataset includes high-quality text while alleviating common artifact patterns introduced by LLMs. Taking advantage of features such as the exploratory user interface, LLM integrations, and prompt templating, was instrumental to speed up the data generation process. The following sections describe the building process in more detail.

3.1 Human Sources

We gather human texts from existing datasets, shown in Table 1. Specifically, we employ the following datasets: OASST-2 (Köpf et al., 2024), Okapi-HS (Lai et al., 2023), WikiLingua (Ladhak et al., 2020), CaSum (de Gibert et al., 2022), XLSum (Hasan et al., 2021), OSCAR (Ortiz Suárez, Sagot, and Romary, 2019), WikiSource,[3] BookSum (Kryscinski et al., 2022), Spanish Books (Ortiz-Fuentes, 2022), ELD,[4] KPT,[5] CaSSA (Gonzalez-Agirre et al., 2024), Amazon Polarity (Zhang, Zhao, and LeCun, 2015), PCSR,[6] B2W-Reviews,[7], TweetLID (Zubiaga et al., 2014), and Wikipedia.[8] Some of these datasets include small quantities of very long texts, in which case we split them into smaller chunks to augment the data.

While for many languages and domains there are existing datasets of human texts, this was not the case for some specific domain-language pairs. For these, we scrap or filter data from other sources. In the literary domain for Catalan and Galician, we scrap literary works from WikiSource. For Galician and Basque news and reviews, first we selected texts in these languages from open-domain corpora such as OSCAR and MC4 (Xue et al., 2021). Then, we classify these texts into *news*, *reviews*, and *other* by using a bi-encoder with `multilingual-e5-large` (Wang et al., 2024a) as backbone. This model is fine-tuned with 10k samples for each label, using XSUM (Narayan, Cohen, and Lapata, 2018) as labeled *news*, Amazon Polarity as *reviews* and Wikipedia and Reddit[9] as *other*. All these English datasets are translated to the target language using `gpt-3.5-turbo-0125`. Additionally, we further filter texts classified as *review*, keeping only those containing the string "!!!", and then removing the pattern, as we found it increases the precision via manual inspection.

At time of writing, there was no publicly available dataset for the how-to domain in Galician, and the filtering pipeline employed for news and reviews yielded small quantities of low quality text, so we opted to not include data for this pair.

3.2 Machine Generated Text

We employ TextMachina to generate MGT from human texts, leveraging its exploration functionality to quickly iterate over prompting techniques and LLMs.

Through manual inspection, we find that most LLMs in the range of 1B to 30B parameters are unable to generate high-quality MGT in Catalan, Basque, and Galician. In most of these cases, the LLMs either confusingly switch to Spanish or Portuguese, or generate objection answers in the form "*I'm sorry, but I do not understand or speak that language*". After a preliminary evaluation of the most representative LLMs, we opt to generate text with `gpt-3.5-turbo-instruct` (Ouyang et al., 2022), `gpt-4` (Achiam et al., 2023), `LLaMa-2-70b-chat-hf` (Touvron et al., 2023), `Mixtral-8x7B-Instruct-v0.1` (Jiang et al., 2024), `cohere.command-text-v14`,[10] and `ai21.j2-ultra-v1`.[11]

Having selected the LLMs, we explore various prompts and decoding parameters for each combination of language, domain, and LLM. Thus, we ensure that the generated texts are of high quality, diverse, and closely

[3] https://ca.wikisource.org/ https://gl.wikisource.org/

[4] https://huggingface.co/datasets/Lam-ia/Euskal-liburu-dataseta

[5] https://www.kaggle.com/datasets/rtatman/brazilian-portuguese-literature-corpus

[6] https://huggingface.co/datasets/beltrewilton/punta-cana-spanish-reviews

[7] https://github.com/americanas-tech/b2w-reviews01

[8] https://www.wikipedia.org/

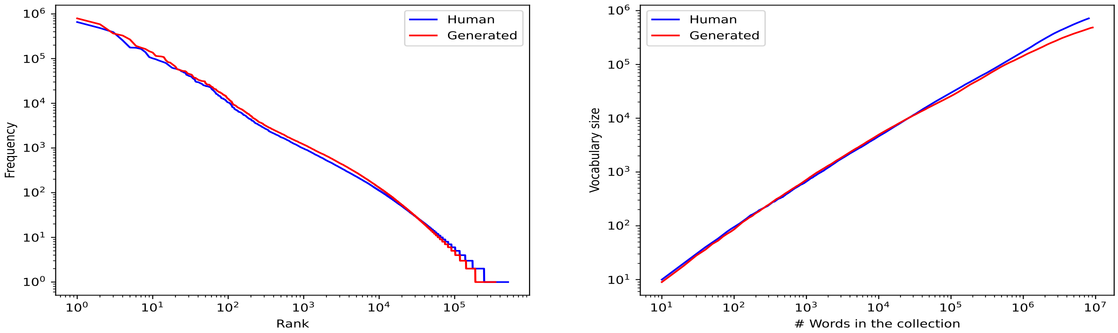
[9] https://huggingface.co/datasets/SophieTr/reddit_clean

[10] https://docs.cohere.com/docs/amazon-bedrock

[11] https://docs.ai21.com/docs/jurassic-2-models

Label		🇬🇧	🇪🇸			🇵🇹
Chat	OASST-2	OASST-2	OASST-2	OASST-2	OASST-2	OASST-2
How-to	Okapi-HS	WikiLingua	Okapi-HS	Okapi-HS	-	Okapi-HS
News	CaSum	XLSum	XLSum	OSCAR†	OSCAR†	XLSum
Literary	WikiSource‡	BookSum	Spanish Books	ELD	WikiSource‡	KPT
Reviews	CaSSA	Amazon Polarity	PCSR	OSCAR†	OSCAR†	B2W-Reviews
Tweets	TweetLID	TweetLID	TweetLID	TweetLID	TweetLID	TweetLID
Wikipedia	Wikipedia	Wikipedia	Wikipedia	Wikipedia	Wikipedia	Wikipedia

Table 1: Human-authored sources for the IberAuTexTification dataset. Cells marked with † are filtered via domain classification, and those marked with ‡ are scraped from the web.



(a) Zipf's law.

(b) Heaps' law.

Figure 1: Zipf's and Heap's laws on human and generated texts.

resemble human texts in content and structure. To support this, Figure 1 depicts Zipf's and Heaps' laws for human and generated texts in the dataset. There is both a large similarity in terms of frequency across word ranks and in terms of vocabulary diversity across word sizes.

3.3 Post-processing

Through the exploration functionality of TextMachina, we find discrepancies between the structure of MGT and human texts that could bias the models to learn spurious correlations.

By default, TextMachina applies a generic post-processing pipeline to deal with these issues, however, we further process all the texts with domain-specific transformations to reduce deviations between MGT and human texts' structure. For the how-to domain, we delete enumerations since human texts rarely enumerate steps. In tweets, we remove emojis and replaced generic user mentions like "user" by random users from the datasets. Regarding literary texts, we remove dialogue symbols like "—" since LLMs are not prone to generate them. For the same reason, we eliminate all the citations from Wikipedia articles.

Furthermore, we (i) remove all the texts with less than 10 words, (ii) replace newlines, tabular, and repeated whitespaces by a single whitespace, and (iii) lowercase and capitalize the texts to ensure both human and generated texts begin with a capital letter.

3.4 Splits

The datasets of both subtasks are split by domain. The training sets contain texts from literary, news, reviews, tweets, and wikipedia domains, while the test sets comprise texts from how-to and chat domains. Thus, we aim to evaluate the generalization capabilities to detect and attribute text on unseen domains.

All the dimensions, namely labels, languages, and LLMs, are generally balanced in the training and test sets of both subtasks, with some exceptions in specific language-domain pairs. Table 2 shows the statistics of the whole IberAuTexTification dataset before splitting training and test sets. For Subtask 2, human texts are removed from the splits, since attribution is done only on MGT. Notably, in this task we anonymize the labels to not disclose the LLMs to the participants during the training phase.

4 Task Organization

The IberAuTexTification shared task was run in two phases:

Dom.	Aut.	🟨	🇬🇧	🇪🇸	🟥	⬜	🇵🇹	Tot.
💬	🤖	2.6	2.8	2.7	1.9	1.4	2.6	14.2
	👩	1.6	2.5	2.5	0.7	0.4	2.0	10.1
HOW TO ?	🤖	2.5	2.9	2.6	2.2	0	2.5	13.0
	👩	2.1	2.6	2.2	1.9	0	2.2	11.3
📚	🤖	1.6	3.2	2.7	2.1	0.7	2.6	13.2
	👩	0.5	2.9	2.7	1.6	0.2	2.4	10.5
📰 NEWS	🤖	2.4	3.1	2.4	2.3	1.8	2.1	14.4
	👩	2.2	2.8	2.3	1.0	1.9	2.3	12.9
👍	🤖	1.8	2.5	1.9	2.2	1.7	2.1	12.5
	👩	0.5	2.9	2.8	0.5	0.4	2.9	10.3
🐦	🤖	2.1	2.5	1.5	1.0	0.8	2.0	10.3
	👩	1.4	1.0	2.0	0.2	0.3	1.5	6.6
🌐	🤖	2.4	2.9	2.4	2.0	1.9	2.1	14.1
	👩	2.3	2.4	2.4	1.7	2.6	2.3	14.0
Tot.		26.7	37.9	33.9	22.1	14.8	32.4	168.1

Table 2: Summarized statistics of the IberAutextification dataset. Sizes in thousands. **Dom.** Domain, **Aut.** Author, **Tot.** Total.

Training phase. The training sets of both subtasks were released in March 22nd for the participants to train their models. Together with the datasets, we released a Github Repository[12] to run the baselines, evaluate predictions, and check the submission format. Thus, participants are able to work on the same environment in which their submissions are evaluated. Since no development set was provided, participants must build their own development sets, ideally resembling the test setting with two unseen domains. Submissions are limited to a maximum of three runs per subtask, with some restrictions. Given that IberAuTexTification datasets have been compiled from publicly available sources, it is necessary to ensure participants do not use data from test partitions as part of their training. They are only allowed to use text derived from the training data, that is, data augmentation, further self-supervised pre-training, or any other technique involving the usage of additional text must be done only from text of the training data.

Test phase. The test sets of both subtasks, comprised of texts from chat and how-to domains, were released in April 21st. Participants were given two months to label the test sets with their best approaches and send their submissions until the deadline in June 10th. After the competition ended, we released the complete dataset in the HuggingFace Hub,[13] with all the information per sample including domain, prompt, language, label, and unanonymized LLM.

5 *Systems*

A total of 21 teams participated in the Subtask 1, submitting 54 runs. In Subtask 2, the participation was smaller, with a total of 7 teams and 14 runs. In this section, we describe the most representative submissions along with our chosen baselines.

5.1 Baselines

We consider various baselines for each subtask, attempting to cover statistical and neural approaches. Specifically, we evaluate a random baseline (*Random*), a bag-of-word and bag-of-character based classifier (*LR+BOW*), a fine-tuned transformer (*Transformer*), and LLMixtic (Sarvazyan, González, and Franco-Salvador, 2024a), the best system of SemEval 2024 Task 8 - MGT Detection (Wang et al., 2024b) (*LLMixtic*). Namely, these consist of the following:

Random. A uniform dummy classifier offered by scikit-learn (Pedregosa et al., 2011).

LR+BOW. We extract bag-of-unigrams and bigrams at word level, and bag-of-characters with 2 to 6-grams, only keeping the 5,000 most frequent n-grams at word and character level. These features are concatenated, normalized using z-score, and fed to a logistic regression classifier with default parameters offered by scikit-learn.

Transformer. We leverage the HuggingFace ecosystem (Wolf et al., 2020) to fine-tune `mdeberta-v3-base` (He, Gao, and Chen, 2022) for 5 epochs, with a batch size of 4, a learning rate of 5e-5, and mixed precision training.

LLMixtic. We extract token-level probabilistic features from LLaMA-2 (Touvron et al., 2023) models, quantizing them to 4 bits: `Llama-2-7b-hf`, `Llama-2-7b-chat-hf`, `Llama-2-13b-hf`, `Llama-2-13b-chat-hf`. These features were fed into an FFN with a hidden dimension of 128, whose output is passed to a single-layer Transformer (Vaswani et al., 2017) encoder with 4 heads and $d_{FFN} = 64$. The encoder's hidden outputs are mean pooled and fed to a final classification head. The model was trained

[12] `https://github.com/Genaios/IberAuTexTification`

[13] `https://huggingface.co/datasets/Genaios/iberautextification`

for 5 epochs with a batch size of 16, a learning rate of 1e-3, and mixed precision training.

5.2 Submitted Approaches

Following common trends in text classification tasks, most approaches were based on n-gram frequency features, neural approaches, or a combination of both. Regarding n-gram based models, most participants fed word-level TF-IDF features to traditional classifiers such as Logistic Regression, Support Vector Machines, Random Forests, and ensembles of these. Notably, many teams relied on pre-processing methods such as stop-word removal, stemming, and lemmatization prior to obtaining these features.

With respect to neural approaches, most of these fine-tune pre-trained Transformer (Vaswani et al., 2017) models. Notably, the variants most participants opted for were multilingual BERT-based models such as `xlm-roberta-base` (Conneau et al., 2020) and `mdeberta-v3-base`. In some cases, teams opted to fine-tune one model for each language, and at inference time they first detect the language of a given text and feed the corresponding fine-tuned Transformer to obtain the predictions. Other approaches also include extracting feature vectors with fasttext (Joulin et al., 2017) and feeding them to a CNN-based classifier, replicating approaches of previous editions (Przybyla, Duran-Silva, and Egea-Gómez, 2023), or using published techniques (Hans et al., 2024), adapting them to our dataset's multilingual and multi-domain scenario. Moreover, some additional approaches also combined these approaches with graph neural networks.

Notably, the best ranked systems for both subtask are comprised of Transformer models augmented with additional lexical, syntactic and semantic features. Specifically for Subtask 1, the best system was proposed by team jor_isa_uc3m (Fernández García and Segura-Bedmar, 2024). This team fine-tunes three multilingual transformers, `distilbert-base-multilingual-cased` (Sanh, 2019), `mdeberta-v3-base`, and `xlm-roberta-base`, making an ensemble with a logistic regression model as a final classifier. For this, they employ 10% of the training data as a validation set for early stopping, and another 10% to train the logistic regression model. On the other hand, for Subtask 2, team gmc_fosunlp (Guo et al., 2024) fine-tunes an `xlm-roberta-base` model, fusing its embeddings with additional frequency, readability, lexical richness, and punctuation features into a final MLP classifier.

6 *Results and Discussion*

6.1 Evaluation

We evaluate submissions for both subtasks with the Macro-F_1 score, reporting statistical significance via bootstrapping with replacement at a confidence level of $\alpha = 0.95$ with 1,000 resamples.

6.2 Subtask 1: MGT Detection

For the MGT detection subtask we received 54 submissions from 21 different teams. Table 3 presents the final ranking including the baselines.

The best system was proposed by team jor_isa_uc3m, obtaining a Macro-F_1 of 80.5%, an improvement of almost 4 points over second place and of over 10 points over the best baseline. In Figure 2 we further analyse these scores. Figure 2a illustrates rank-ordered Macro-F_1 scores, from where we observe a linear distribution with three outliers ranging from 75 to 80 Macro-F_1. Notably, most submissions score between the best and worst baseline, with only eight teams beating the best baseline. 21 teams scored below the Transformer baseline, and all the teams outperform the random baseline by at least 7 points of Macro-F_1. Moreover, from the precision-recall distributions shown in Figure 2 we observe that systems are more biased toward predicting text as generated (high recall), often incorrectly (low precision), while the opposite applies to the human class. This is in line with what was observed in the previous edition of the shared task.

We carry out a more fine-grained analysis into the submitted predictions to understand their performance in specific domains, languages, labels, and authors. These results are presented in Figure 3.

Domain-wise results. When observing Macro-F_1 scores in Figure 3a, we see that systems generalize better to the how-to domain than to the chat domain, with the former exhibiting a much wider distribution with very high variance, while the latter is a more peaked distribution. This is not surprising, since chat utterances are shorter than how-to articles, making it more difficult to find discriminative patterns.

Language-wise results. From language-wise Macro-F$_1$ scores in Figure 3b, we observe that most languages are distributed in a similar manner. Almost all the boxes overlap, with medians close to 63% Macro-F$_1$ and similar variances. Some interesting cases are Basque with a much higher variance, and Galician, with the lowest median value of approximately 57% Macro-F$_1$. This proves that detecting MGT is more difficult in these two low-resource languages, even when detectors are provided a similar amount of training data for all the languages.

Label-wise results. We observe in Figure 3c that systems are better at correctly identifying generated text, and there is lower dispersion among F$_1$ scores for this class than for the human class. For the human class, half of the systems perform on par to the random baseline. Again, these behaviours hold from the previous edition of this task.

Author-wise results. Presented in Figure 3d which illustrates very small differences in performance when identifying generated text from any of the LLMs. All the box-plots belonging to LLMs overlap, there is no statistically significant differences, suggesting that the difficulty of classifying MGT across these LLMs is very similar. Otherwise, there are likely statistical differences between human and LLM's F$_1$ scores, being the former more difficult to be correctly classified.

In order to understand the difficulty of the subtask, we compute distributions of easy and hard examples. Here, we measure the difficulty of each example in the test set by total percentage of participants that correctly predicted it. We do this analysis at domain, language, label and author level, as shown in Figure 4.

Domain-wise difficulty. Figure 4a illustrates that both chat and how-to are distributed similarly in terms of difficulty, with the chat domain having slightly more hard examples and less easy ones than how-to.

Language-wise difficulty. We find in Figure 4b that all the languages have a large amount of easy texts, with Basque, Catalan, and Portuguese predominating. Also, all the languages show a similar number of hard texts. Most of them comprise the same amount of medium difficulty texts, with English, Spanish, and Galician containing more examples in the middle-easy range.

Rank	Team	Run	Macro-F$_1$
1	jor_isa_uc3m	jor_isa_ens_multi	**80.50**
2	gmc_fosunlp	run1	76.63
3	telescope_team	run2	75.79
4	iimasNLP	s1_gnn_stylo	71.88
5	gmc_fosunlp	run2	71.55
9	baselines	LLMixtic	69.84
12	Achraf	run1	67.84
13	baselines	LR+BOW	67.67
35	Joavpa	run3	61.82
36	baselines	Transformer	61.47
57	RUns	run3	56.08
58	baselines	Random	49.72

Table 3: Truncated ranking of Subtask 1, showing the top-5 systems, the baselines, and the participant before each baseline.

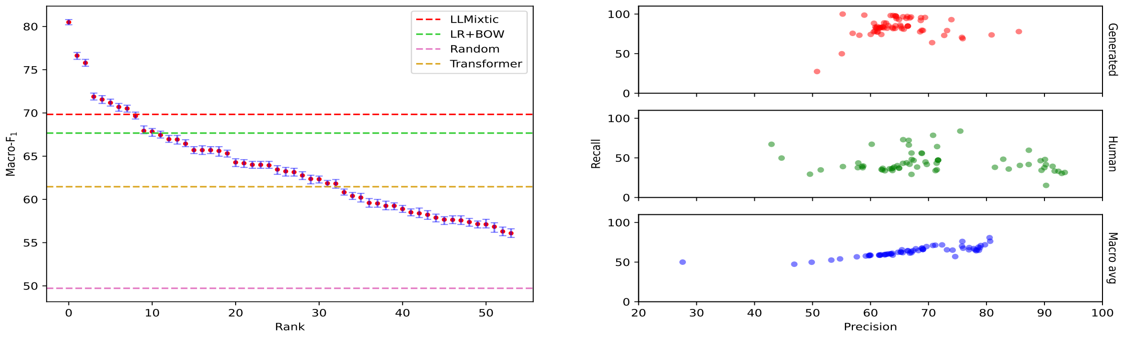
Label-wise difficulty. In Figure 4c, we see that human examples are more scattered across the difficulty range, and there are less examples in the medium-easy range than in the hard-medium one, where generated texts are scarce. This follows from our previous fine-grained analysis, where we found that systems are better at correctly classifying generated text than human-authored text.

Author-wise difficulty. We corroborate our previous author-wise results in Figure 4d, where all the LLMs are similarly distributed, only showing differences with respect to the peak sizes.

6.3 Subtask 2: Model Attribution

In the MGT attribution task we received 14 submissions from 7 different teams. We present the final ranking with the baselines in Table 4.

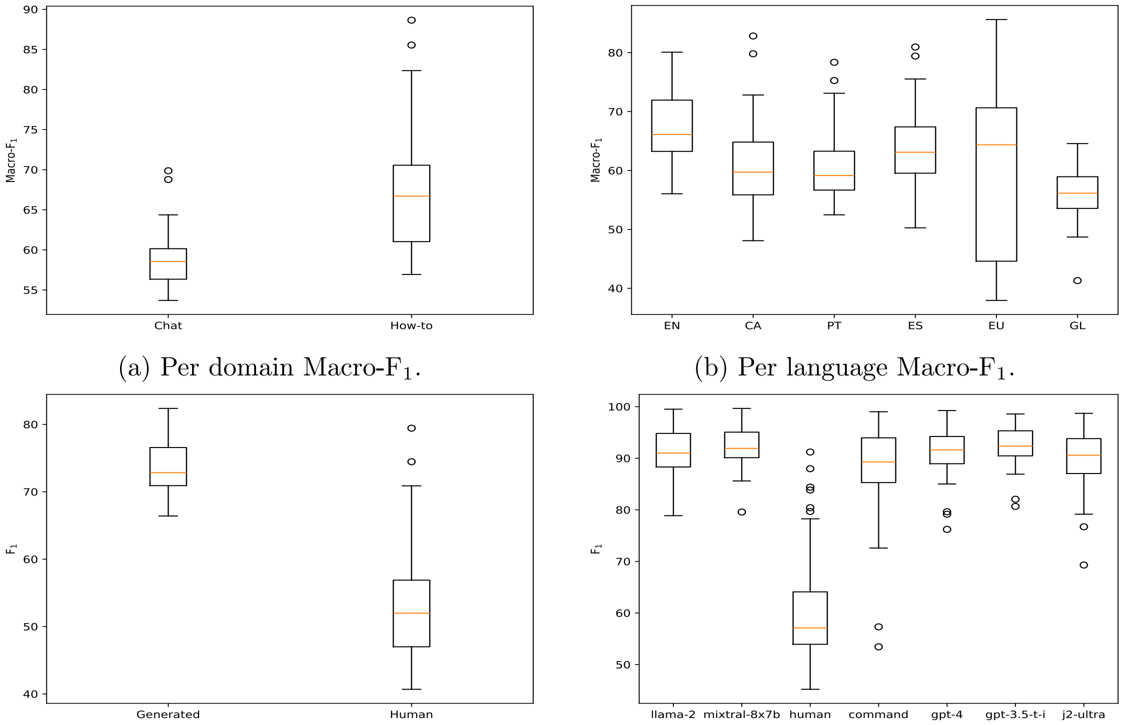
Team gmc_fosunlp proposed the best system, obtaining 52.31% Macro-F$_1$, an improvement of almost 6 points over the best baseline. We present the rank-ordered Macro-F$_1$ scores of the submitted systems in Figure 5. We observe that most submissions score better than the best baseline, while some are between the two worse-performing baselines. Notably, most teams outperform the random baseline by at least 15 points, with only one ranking below it. Eight teams outperform the best baseline but do not deviate from it by more than 6 points of Macro-F$_1$. This smaller difference between the best participants and the best baseline highlights the higher difficulty of this task compared to Subtask 1.



(a) Rank-ordered Macro-F$_1$ scores with error bars. Dotted lines are baselines.

(b) Precision-recall distributions.

Figure 2: Precision, Recall and Macro-F$_1$ plots for Subtask 1.



(a) Per domain Macro-F$_1$.

(b) Per language Macro-F$_1$.

(c) Per label F$_1$.

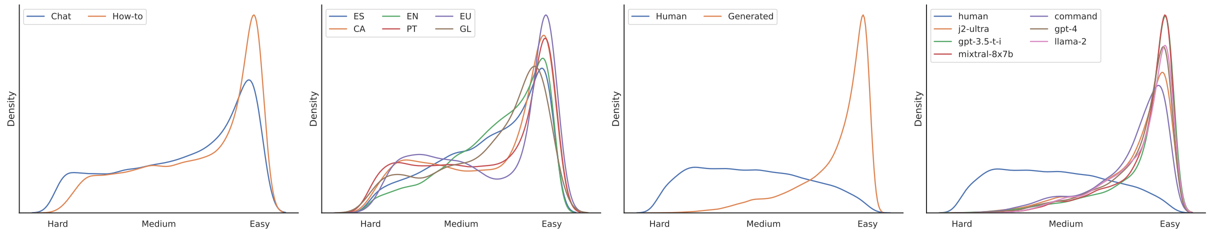
(d) Per author F$_1$.

Figure 3: Fine-grained analysis of Subtask 1 submissions.

In the same manner as for Subtask 1, we carry out a more fine-grained analysis of the submitted predictions in specific domains, languages, and labels. These results are presented in Figure 6.

Domain-wise results. Depicted in Figure 6a, we observe that participant's systems are worse attributing texts to LLMs in the chat domain than in the how-to domain, similarly to Subtask 1 for MGT detection. Both domains exhibit similar distributions, with half of the participants scoring below 40 Macro-F$_1$ in the chat domain, and below 50 Macro-F$_1$ in the how-to domain.

Language-wise results. From Figure 6b, we see that, while there are no statistically significant differences between per-language Macro F$_1$ scores, it is interesting to highlight the special case of Basque, where participants achieve much better performance with respect to the other languages. Other interesting cases include English with much higher variability, and Spanish and Portuguese with very low variability. Additionally, in all the languages we observe median Macro-F$_1$ scores of 40% or better, with many distribu-



(a) Per domain. (b) Per language. (c) Per label. (d) Per author.

Figure 4: Distributions of the number of correct predictions in Subtask 1. The *Hard* label shows where zero participants correctly labeled the text, while the *Easy* label marks texts that were correctly classified by all the systems.

tional outliers dropping as low as 15%.

Label-wise results. We observe in Figure 6c that attributing text to `command` is easier than for other LLMs. Despite their differences in generation capabilities, e.g., `gpt-4` vs `LLaMA-2`, there are no statistical differences in the attribution label-wise F_1 scores. Except for `command`, all the box-plots overlap and are similarly distributed. This is especially true for `Mixtral-8x7b`, `gpt-4`, and `gpt-3.5-turbo-instruct`, with the last two having a more similar median F_1 score.

We carry out an analysis to assess the task difficulty as in Subtask 1, obtaining distributions of easy and hard examples measured by the total percentage of participants that correctly attribute a text to the LLM that authored it. The per-domain, per-language and per-label distributions are illustrated in 7.

Domain-wise difficulty. Figure 7a shows that chat texts are much more difficult to attribute to a specific LLM than how-to texts. On the other hand, these appear to be more uniformly distributed across the difficulty spectrum.

Language-wise difficulty. From 7b we corroborate that Basque exhibits a much larger quantity of easy examples. The remaining languages include many more difficult examples than easy ones. However, all the languages have a similar number of middle-difficulty texts, with most of these leaning toward hard-medium difficulty.

Label-wise difficulty. Illustrated in Figure 7c, we find that `j2-ultra` and `command` mainly generate easily-attributable text. `Mixtral-8x7b` and `LLaMA-2` have more texts of medium difficulty, while `gpt-4` and `gpt-3.5-turbo-instruct` generate text that is hard to attribute to them correctly. This is especially interesting considering they are different LLM versions of the same provider.

Rank	Team	Run	Macro-F_1
1	gmc_fosunlp	run1	**52.31**
2	iimasNLP	s2_llm3_stylo	51.73
3	Drocks	run2	50.75
4	Drocks	run1	50.30
5	iimasNLP	s2_llm2_gnn_stylo	49.58
9	baselines	Transformer	46.50
10	baselines	LLMixtic	45.55
11	baselines	LR+BOW	42.16
16	ahoumaine	run2	30.34
17	baselines	Random	16.82

Table 4: Truncated ranking of Subtask 2, showing the top-5 systems, the baselines, and the participant before each baseline.

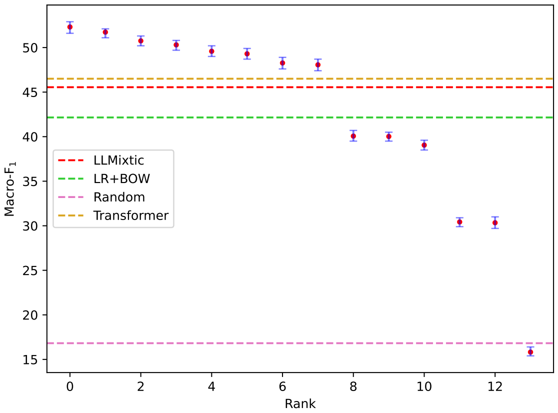
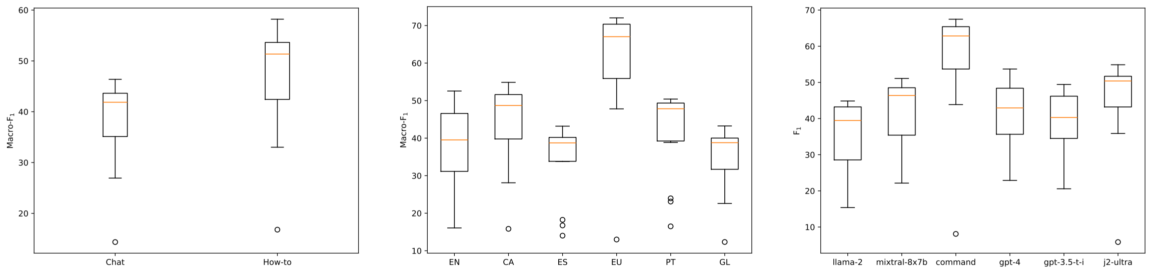


Figure 5: Rank-ordered Macro-F_1 scores with error bars for Subtask 2. Dotted lines are baselines.

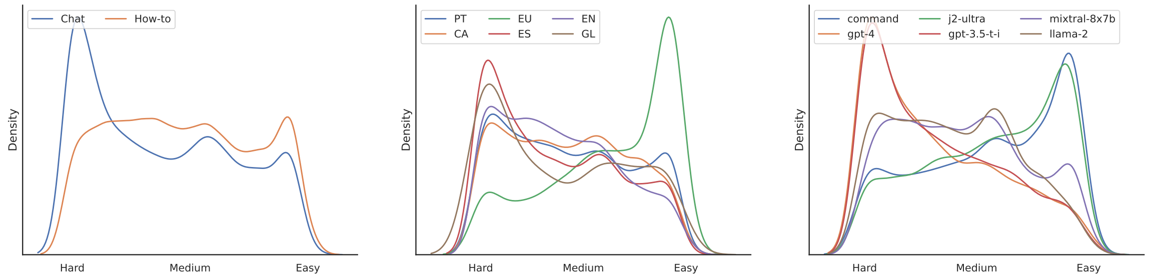
7 *Conclusions and Future Work*

In this paper we present the IberAuTexTification shared task at IberLEF 2024, where we study MGT detection and attribution in a multilingual, multi-model, and multi-domain setting, focusing on (i) seven domains, (ii) the six main languages spoken in



(a) Per domain Macro-F$_1$. (b) Per language Macro-F$_1$. (c) Per label F$_1$.

Figure 6: Fine-grained analysis of Subtask 2 submissions.



(a) Per domain. (b) Per language. (c) Per label.

Figure 7: Distributions of the number of correct predictions in Subtask 2. The *Hard* label shows where zero participants correctly labeled the text, while the *Easy* label marks texts that were correctly classified by all the systems.

the Iberian peninsula, and (iii) six of the most representative LLMs. The IberAuTexTification dataset has been generated through TextMachina, consisting of 168,000 human-written and high-quality, diverse, human-like, machine-generated texts.

The shared task received a significant amount of participation, especially for Subtask 1. A total of 21 teams participated in Subtask 1 with 54 runs, while 7 teams submitted 14 runs to the Subtask 2. The participating systems relied on a wide variety of approaches, with a strong trend towards the use of Transformer models. Ensembles of fine-tuned transformers led to the best results on Subtask 1, while the best results in Subtask 2 were reached by combining n-gram frequency features with a fine-tuned transformer.

Our findings show that many teams struggle with MGT detection and, particularly, with attribution, under cross-domain settings, especially when involving low-resource languages. In Subtask 1, most submissions do not outperform a logistic regression baseline, though some do significantly better. Additionally, it seems easier to generalize to how-to articles than to chat utterances with the provided training data. We also found similar difficulties in detecting MGT from any of the considered LLMs. Unsurprisingly, this task is more challenging in low-resource languages like Galician or Basque. In Subtask 2, most of the models outperform the best baseline. It is also easier to generalize to how-to articles than to chat utterances. Besides, despite the LLM's differences in generation capabilities, the participants' F$_1$ scores across LLMs are very similar. Interestingly, attributing text in Basque appears easier than in other languages.

As future work, we plan to extend both the task and TextMachina to other modalities like image or speech and to update the dataset with more LLMs, domains, and languages to encourage the development of more robust and generalizable systems.

References

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Antoun, W., B. Sagot, and D. Seddah. 2024. From text to source: Results in detecting large language model-generated content. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7531–7543, Torino, Italia, May. ELRA and ICCL.

Bakhtin, A., S. Gross, M. Ott, Y. Deng, M. Ranzato, and A. Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351.*

Chiruzzo, L., S. M. Jiménez-Zafra, and F. Rangel. 2024. Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org.*

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

de Gibert, O., K. Kharitonova, B. C. Figueras, J. Armengol-Estapé, and M. Melero. 2022. Sequence-to-sequence resources for catalan. *arXiv preprint arXiv:2202.06871.*

Eloundou, T., S. Manning, P. Mishkin, and D. Rock. 2024. Gpts are gpts: Labor market impact potential of llms. *Science*, 384(6702):1306–1308.

Fernández García, J. and I. Segura-Bedmar. 2024. Human after all: Using transformers based models to identify automatically generated text. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024). CEUR Workshop Proceedings, CEUR-WS, Valladolid, Spain.*

Fivez, P., W. Daelemans, T. Van de Cruys, Y. Kashnitsky, S. Chamezopoulos, H. Mohammadi, A. Giachanou, A. Bagheri, W. Poelman, J. Vladika, et al. 2024. The clin33 shared task on the detection of text generated by large language models. *Computational Linguistics in the Netherlands Journal*, 13:233–259.

Gonzalez-Agirre, A., M. Marimon, C. Rodriguez-Penagos, J. Aula-Blasco, I. Baucells, C. Armentano-Oller, J. Palomar-Giner, B. Kulebi, and M. Villegas. 2024. Building a data infrastructure for a mid-resource language: The case of Catalan. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2556–2566, Torino, Italia, May. ELRA and ICCL.

Guo, M., Z. Han, X. Wang, and J. Peng. 2024. Multidimensional text feature analysis: Unveiling the veil of automatically generated text. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024). CEUR Workshop Proceedings, CEUR-WS, Valladolid, Spain.*

Hans, A., A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, and T. Goldstein. 2024. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *Forty-first International Conference on Machine Learning.*

Hasan, T., A. Bhattacharjee, M. S. Islam, K. Mubasshir, Y.-F. Li, Y.-B. Kang, M. S. Rahman, and R. Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online, August. Association for Computational Linguistics.

He, P., J. Gao, and W. Chen. 2022. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations.*

Henderson, P., X. Li, D. Jurafsky, T. Hashimoto, M. A. Lemley, and P. Liang. 2023. Foundation models and fair use. *Journal of Machine Learning Research*, 24(400):1–79.

Ippolito, D., D. Duckworth, C. Callison-Burch, and D. Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online, July. Association for Computational Linguistics.

Jiang, A. Q., A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088.*

Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov. 2017. Bag of tricks for efficient text classification. In M. Lapata, P. Blunsom, and A. Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April. Association for Computational Linguistics.

Kasneci, E., K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, page 102274.

Köpf, A., Y. Kilcher, D. von Rütte, S. Anagnostidis, Z. R. Tam, K. Stevens, A. Barhoum, D. Nguyen, O. Stanley, R. Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Kryscinski, W., N. Rajani, D. Agarwal, C. Xiong, and D. Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Ladhak, F., E. Durmus, C. Cardie, and K. McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online, November. Association for Computational Linguistics.

Lai, V., C. Nguyen, N. Ngo, T. Nguyen, F. Dernoncourt, R. Rossi, and T. Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore, December. Association for Computational Linguistics.

Liu, Y., T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, et al. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852.*

Mitchell, E., Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *International Conference on Machine Learning.*

Molla, D., H. Zhan, X. He, and Q. Xu. 2023. Overview of the 2023 alta shared task: Discriminate between human-written and machine-generated text. In *Proceedings of the 2023 Australasian Language Technology Association Workshop (ALTA 2023).*

Narayan, S., S. B. Cohen, and M. Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, October-November. Association for Computational Linguistics.

Nasr, M., N. Carlini, J. Hayase, M. Jagielski, A. F. Cooper, D. Ippolito, C. A. Choquette-Choo, E. Wallace, F. Tramèr, and K. Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035.*

Ortiz-Fuentes, J. 2022. Crawled spanish books.

Ortiz Suárez, P. J., B. Sagot, and L. Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Przybyla, P., N. Duran-Silva, and S. Egea-Gómez. 2023. I've seen things you machines wouldn't believe: Measuring content predictability to identify automatically-generated text. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023). CEUR Workshop Proceedings, CEUR-WS, Jaén, Spain.*

Sanh, V. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019).*

Sarvazyan, A. M., J. Á. González, and M. Franco-Salvador. 2024a. Genaios at SemEval-2024 task 8: Detecting machine-generated text by mixing language model probabilistic features. In A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, editors, *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 101–107, Mexico City, Mexico, June. Association for Computational Linguistics.

Sarvazyan, A. M., J. Á. González, and M. Franco-Salvador. 2024b. Textmachina: Seamless generation of machine-generated text datasets. *arXiv preprint arXiv:2401.03946.*

Sarvazyan, A. M., J. Á. González, M. Franco-Salvador, F. Rangel, B. Chulvi, and P. Rosso. 2023a. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *Sociedad Española de Procesamiento del Languaje Natural (SEPLN)*, 71:275–288.

Sarvazyan, A. M., J. A. González, M. Franco-Salvador, and P. Rosso. 2023b. Supervised machine-generated text detectors: Family and scale matters. In *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization.* Springer International Publishing.

Team, G., R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805.*

Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288.*

Uchendu, A., T. Le, K. Shu, and D. Lee. 2020. Authorship attribution for neural text generation. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online, November. Association for Computational Linguistics.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, L., N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei. 2024a. Multilingual e5 text embeddings: A technical report.

Wang, Y., J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, O. Mohammed Afzal, T. Mahmoud, G. Puccetti, and T. Arnold. 2024b. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, editors, *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico, June. Association for Computational Linguistics.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Xue, L., N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June. Association for Computational Linguistics.

Zellers, R., A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Zhang, Q., C. Gao, D. Chen, Y. Huang, Y. Huang, Z. Sun, S. Zhang, W. Li, Z. Fu, Y. Wan, and L. Sun. 2024. LLM-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In K. Duh, H. Gomez, and S. Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 409–436, Mexico City, Mexico, June. Association for Computational Linguistics.

Zhang, X., J. Zhao, and Y. LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zubiaga, A., I. San Vicente, P. Gamallo, N. Aranberri, and A. Ezeiza. 2014. Overview of tweetlid: Tweet language identification at sepln 2014. In *Procesamiento del Lenguaje Natural*.