Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 275-288 recibido 27-06-2023 revisado 07-07-2023 aceptado 16-07-2023

Overview of AuTexTification at IberLEF 2023: Detection and Attribution of Machine-Generated Text in Multiple Domains

Resumen de AuTexTification en IberLEF 2023: Detección y Atribución de Texto Generado Automáticamente en Múltiples Dominios

Areg Mikael Sarvazyan,[1] **José Ángel González,**[1] **Marc Franco-Salvador,**[1] **Francisco Rangel**,[1] **Berta Chulvi**,[2] **Paolo Rosso**[2]
[1]Symanto Research, Valencia, Spain
[2]Universitat Politècnica de València, Valencia, Spain
{areg.sarvazyan, jose.gonzalez, marc.franco, francisco.rangel}@symanto.com
berta.chulvi@upv.es, prosso@dsic.upv.es

Abstract: This paper presents the overview of the AuTexTification shared task as part of the IberLEF 2023 Workshop in Iberian Languages Evaluation Forum, within the framework of the SEPLN 2023 conference. AuTexTification consists of two subtasks: for Subtask 1, participants had to determine whether a text is human-authored or has been generated by a large language model. For Subtask 2, participants had to attribute a machine-generated text to one of six different text generation models. Our AuTexTification 2023 dataset contains more than 160.000 texts across two languages (English and Spanish) and five domains (tweets, reviews, news, legal, and how-to articles). A total of 114 teams signed up to participate, of which 36 sent 175 runs, and 20 of them sent their working notes. In this overview, we present the AuTexTification dataset and task, the submitted participating systems, and the results.
Keywords: Machine-Generated Text, Large Language Models, Generalization, AuTexTification.

Resumen: Este artículo presenta un resumen de la tarea AuTexTification como parte del workshop IberLEF 2023 sobre el Iberian Languages Evaluation Forum, en el marco de la conferencia SEPLN 2023. AuTexTification consta de dos subtareas: en la Subtarea 1, los participantes tuvieron que determinar si un texto fue escrito por un humano o generado por un modelo de lenguaje masivo. Para la Subtarea 2, los participantes debían atribuir un texto generado automáticamente a uno de seis modelos de generación de texto diferentes. El conjunto de datos AuTexTification contiene más de 160.000 textos en dos idiomas (inglés y español) y cinco dominios (tweets, reseñas, noticias, legislación y artículos instructivos). Un total de 114 equipos se inscribieron para participar, de los cuales 36 enviaron 175 resultados y 20 de ellos enviaron artículos. En este artículo, presentamos el conjunto de datos y la tarea AuTexTification, los sistemas enviados por los participantes y sus resultados.
Palabras clave: Texto Generado por Máquina, Modelos de Lenguaje Masivos, Generalización, AuTexTification.

1 Introduction

Current developments in Large Language Models (LLMs) have strongly improved the quality of Machine-Generated Text (MGT). Their latest surge in popularity through services such as ChatGPT,[1] and large-scale democratization efforts to broaden the public's access to large models (Scao et al., 2022; Touvron et al., 2023; Wolf et al., 2020; Seger et al., 2023), have made it easier for non-technical people to interact with and use these models for various interesting applications (Eloundou et al., 2023; Liu et al., 2023).

However, these advances have also lowered

[1]https://tinyurl.com/reuters-chatgpt

ISSN 1135-5948 DOI 10.26342/2023-71-21

Human Mod.	Model adaptation	
	Pre-trained	**Fine-tuned**
No	Full accessibility Few comp. resources Massive scale Medium quality	Technical accessibility Large comp. resources Massive scale High quality
Yes	High accessibility Few comp. & human resources Small scale High quality	Technical accessibility Large comp. & human resources Small scale High quality

Table 1: Types of MGT. The AuTexTification 2023 Shared Task focuses on generations from pre-trained models with no human modification. We cover the most accessible approach, involving little computational and human resources and can be used massively.

the barrier of entry for users to generate high-quality, multi-style and multi-domain text in a massive scale. This means that motivated malicious users could easily generate massive quantities of text without the need of large computational resources, technical knowledge, or human intervention (see Table 1). Supporting this concern, recent research suggest that disinformation generated with state-of-the-art LLMs is more credible than the one generated by humans (Spitale, Biller-Andorno, and Germani, 2023), thus showing the difficulty for humans to distinguish between MGT and human-authored text.

As expected, the aforementioned advancements have also promoted discussions in ethical AI (Widder et al., 2022) as well as model, data and training regulations,[2] and new licenses (Benjamin et al., 2019; Contractor et al., 2022). Content moderation due to AI democratization, and the need for regulations, are strong motivators for researchers to ensure a responsible use of LLMs and their generations. A promising research line to carry this out involves identifying MGT, while also attributing it to specific text generation models to learn about the specific actors behind an MGT from a forensics viewpoint.

There have been many efforts to detect MGT, including zero-shot approaches (Mitchell et al., 2023; Zellers et al., 2019a), supervised systems (Ippolito et al., 2020; Uchendu et al., 2020; Maronikolakis, Schütze, and Stevenson, 2021), and evaluation campaigns (Kashnitsky et al., 2022; Shamardina et al., 2022). While it has been found that in-domain MGT detection with supervised approaches is easy (Bakhtin et al., 2019), most of the works often overlooked that MGT detection systems would be applied to a broad variety of domains, writing styles, and generation models. Therefore, there is a need to evaluate the generalization of MGT detectors through a more realistic lens. In this regard, some works have studied generalization across model families and scales (Sarvazyan et al., 2023), however, the generalization to new domains is still underexplored.

In this context, we present the AuTexTification (**Au**tomated **Tex**t Iden**Tification**) task. This shared task is proposed to study: (i) the automatic detection of MGT, (ii) the generalization capabilities of MGT detectors to new domains, and (iii) the feasibility of fine-grained MGT attribution to one of many generation models. Furthermore, we automatically collect a multi-domain annotated dataset of human-authored text and MGT generated by various LLMs, which is a valuable resource for exploratory linguistic analysis of machine-generated and human-authored texts. To our knowledge, AuTexTification is the first shared task to study both MGT detection and attribution in a multi-domain setting for English and Spanish, while also focusing on generalization of MGT detectors to new domains.

2 Task Description

The AuTexTification 2023 Shared Task includes two subtasks in English and Spanish in five different domains.

Subtask 1: MGT Detection. This subtask consists in distinguishing between human and generated text. It is framed as a binary classification task of human text (Hum) and MGT (Gen), where text from three domains is included in the training set, and submissions are evaluated in two unseen ones. This way, we aim to study the MGT detectors' cross-domain generalization capabilities.

Subtask 2: MGT Attribution. In this subtask, participants must attribute MGT to the model that generated it, out of six models. Thus, Subtask 2 is framed as a six-class classification task, where we strive to study the feasibility of fine-grained attribution. Differently to Subtask 1, the training and test splits include all five domains.

[2]European Commission, Proposal for a Regulation of the European Parliament `https://tinyurl.com/EURAIAct`

	English	Spanish
Legal	*MultiEURLEX*	*MultiEURLEX*
News	*XSUM*	*MLSUM & XLSUM*
Reviews	*Amazon Reviews*	*COAR & COAH*
Tweets	*TSATC*	*XLM-Tweets & TSD*
How-to	*WikiLingua*	*WikiLingua*

Table 2: Human-authored source datasets for the AuTexTification 2023 dataset.

3 Dataset

The AuTexTification dataset consists of texts written by humans and LLMs in five domains: tweets, reviews, how-to articles, news and legal documents. These domains were chosen to encompass a range of writing styles, from more structured and formal to less structured and more informal. We collected human texts from publicly available datasets, namely: *MultiEURLEX* (Chalkidis, Fergadiotis, and Androutsopoulos, 2021), *XSUM* (Narayan, Cohen, and Lapata, 2018), *XLSUM* (Hasan et al., 2021), *MLSUM* (Scialom et al., 2020), *Amazon Reviews* (McAuley and Leskovec, 2013), *WikiLingua* (Ladhak et al., 2020), *COAR & COAH* (González et al., 2014), *XLM-Tweets* (Barbieri, Espinosa Anke, and Camacho-Collados, 2022), *TSATC* (Naji, 2012), and *TSD* (Leis et al., 2019). Table 2 groups these datasets per domain and language.

The MGT was generated from the human texts by using three *BLOOM* models (Scao et al., 2022), *BLOOM-1B7*,[3] *BLOOM-3B*,[4] and *BLOOM-7B1*;[5] as well as three *GPT-3* models (Brown et al., 2020; Ouyang et al., 2022): *babbage*, *curie*, and *text-davinci-003*, with 1b, 6.7b and 175b parameter scales respectively. Our motivation behind using these models were fourfold: (i) both *BLOOM* and *GPT-3* show great capabilities in multiple languages, (ii) *BLOOM* models' usage is not as restricted via licensing (as opposed to other popular models such as *LLaMA* (Touvron et al., 2023) or *OPT* (Zhang et al., 2022a)), (iii) *GPT-3* has been one of the most popular and best performing language models until recently,[6] and (iv) we aimed to cover a broad spectra of model families and scales. While we were hoping to include *BLOOM-175B* generations too, this was not possible due to the lack of public APIs.

[3] `https://tinyurl.com/bloom-1b7`

[4] `https://tinyurl.com/bloom-3b`

[5] `https://tinyurl.com/bloom7b`

[6] *GPT-3.5-turbo* and *GPT-4* were not released at time of compiling our dataset.

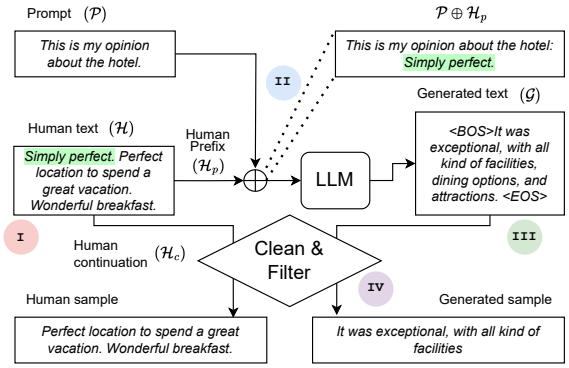


Figure 1: Data gathering process.

We manually tuned the decoding parameters to obtain MGT that appears realistic through subjective evaluations carried out by two of the authors. We found that with nucleus sampling (Holtzman et al., 2020), using a top-p of 0.9 and a temperature of 0.7, the models generated texts of higher quality. The maximum number of completion tokens was manually selected for each domain to be similar to the median token-length of the human texts: 20 tokens for tweets, 70 for reviews, and 100 for news, legal, and how-to articles.

3.1 Gathering process

We aim to build a dataset of human and generated texts that share the same prefix. For instance, given a human text "*Today it's 20 degrees. It is sunny in Valencia.*", we could use "*It is sunny in Valencia.*" as human text, and generate a continuation by prompting an LLM with "*Today it's 20 degrees.*". In this manner, both generated and human texts are plausible continuations of the same prefix and they can be compared fairly in terms of topics and domains. To build the dataset in this way, we opted for a data gathering process consisting in the steps depicted in Figure 1, namely (i) gathering human data, (ii) preparing the inputs for LLMs, (iii) generating MGTs, and (iv) cleaning and filtering the resulting texts.

We first gather a set of human-authored texts $\mathcal{H}$ from the source datasets for each domain and language. We manually analyze and define extraction schemes for splitting $\mathcal{H}$ into prefixes $\mathcal{H}_p$ and continuations $\mathcal{H}_c$ such that $\mathcal{H} = \mathcal{H}_p \oplus \mathcal{H}_c$. In some domains and source datasets, we also define prompts $\mathcal{P}$ to prevent the generation models from gen-

erating topic-inconsistent texts, e.g., guiding models to generate hotel reviews instead of car reviews when using a prefix from the *COAH* dataset, made up of hotel reviews. Afterwards, the prompts and prefixes $\mathcal{P} \oplus \mathcal{H}_p$ are fed into each LLM to obtain one resulting generation per prompt and prefix. We refer to the set of generations as $\mathcal{G}$. Texts from both $\mathcal{H}_c$ and $\mathcal{G}$ are fed into a text cleaning pipeline that removes duplicated spaces, multiple line breaks, and special symbols. Additionally, we ensure that the human continuation and generation obtained from the same prefix have roughly the same token-lengths by truncating to the minimum length of the two texts, thus removing token-length bias. Then, we apply a set of language identification filters: *langdetect*,[7] *SpaCy FastLang*,[8] and *fastText* (Joulin et al., 2017). If one of these filters finds a text to be not in Spanish or English, the text is removed from our dataset.

To obtain the dataset for Subtask 1, we sample a subset of $\mathcal{H}_c$ labeled as HUM and a subset of $\mathcal{G}$ labeled with GEN. The dataset was then split into training and test sets for a cross-domain scenario: tweets, how-to articles and legal documents were included in the training set, while reviews and news data comprised the test set. To compile the dataset for Subtask 2, we only sample texts from $\mathcal{G}$, labeling each text with the LLM's name that generated it. The dataset is randomly split into training and test sets following 80%-20% proportions. All the five domains are included in both training and test splits. The released version of the dataset for Subtask 2 includes anonymized model labels to remove bias toward particular models or model families in participating submissions.

The statistics of each subtask's contents per domain, class, and language are presented in Table 3. In both subtasks, both languages contain similar amounts of texts, and the domains and classes are balanced in both splits. This way we guarantee that our analysis is fair by ensuring that every dimension is balanced. Besides, we checked that the generated texts follow the Zipf and Heap's empirical laws, thus ensuring a high quality of the dataset.[9]

[7] `https://tinyurl.com/langdetect`
[8] `https://tinyurl.com/fastlang`
[9] See `https://tinyurl.com/overview-datasets`

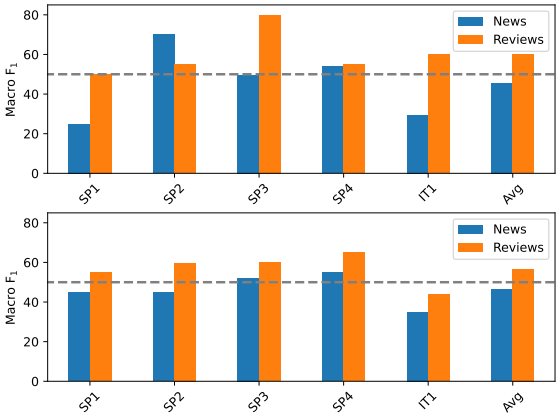


Figure 2: Human performance in English (top) and Spanish (bottom). The grey dotted line is the random baseline.

3.2 Human Assessment

We performed a small-scale study to assess the difficulty of the Subtask 1 for human annotators. The study consisted in asking human annotators to classify texts as human or generated.[10] Five annotators were involved: four Spanish native speakers (SP) and one Italian native speaker (IT). All of them were men between the ages of 25 and 35, with C1-C2 proficiency level in English. From these annotators, SP1 and SP4 are familiar with generated text (they created the dataset and analysed hundreds of examples), while the others were exposed to the task for the first time.

We provided the same 40 texts to each annotator, drawn from the test set of the Subtask 1 both for English and Spanish. The texts were balanced in terms of classes and domains: 20 texts were generated by LLMs and 20 were written by a human, half of them were news and the other half were reviews. The generated texts were only obtained from *BLOOM* models: 6 texts from *BLOOM-1b7*, 6 texts from *BLOOM-3b1*, and 8 texts from *BLOOM-7b1*. Figure 2 shows the Macro-F_1 score of each annotator in each domain.

For both languages, the average annotator performance is very similar, most annotators are close to the random baseline. Regarding the domains, it seems more difficult for humans to distinguish between human-authored and machine-generated news rather than reviews. Most of the annotators per-

[10] The annotation interface and instructions are available at `https://tinyurl.com/colab-annotation`

		Subtask 1			Subtask 2						
					BLOOM			GPT			
		GEN	HUM	Σ	1b7	3b	7b1	1b	6b7	175b	Σ
Spanish	**Legal**	4,846	4,358	9,204	640	665	712	919	942	919	4,797
	News	5,514	5,223	10,737	839	860	881	972	978	987	5,517
	Reviews	5,695	3,697	9,392	952	962	935	945	941	947	5,682
	Tweets	5,739	5,634	11,373	967	965	965	928	930	964	5,719
	How-to	5,690	5,795	11,485	894	929	960	970	983	966	5,702
	Total	27,484	24,707	52,191	4,292	4,381	4,453	4,734	4,774	4,783	27,417
English	**Legal**	5,124	5,244	10,368	809	779	832	890	887	927	5,124
	News	5,464	5,464	10,928	747	854	906	983	984	984	5,458
	Reviews	5,726	5,178	10,904	944	946	939	977	974	972	5,752
	Tweets	5,813	5,884	11,697	987	968	980	951	963	969	5,818
	How-to	5,862	5,918	11,780	962	976	982	993	993	963	5,869
	Total	27,989	27,688	55,677	4,449	4,523	4,639	4,794	4,801	4,815	28,021

Table 3: Number of samples per domain, class, and language in both subtasks.

form worse than the random baseline distinguishing texts from the news domain. On the contrary, humans are typically better than the random baseline in the reviews domain, especially in English.

Language proficiency seems to play a role. IT1 shows better performance in English than in Spanish, where he is not proficient. Despite how SP1 and SP4 are familiar with generated texts, there seems to be no significant difference between them and other annotators.

The human annotators did not follow any systematic pattern to detect MGT. For reviews, some mentioned that the generated reviews seemed generic, describing many general aspects with short sentences. In contrast, human reviews focused on few and more concrete aspects.

4 Systems and Results

In this section, we briefly introduce the participants' systems, describe the baselines and evaluation metrics, and study the results of the shared task.

4.1 Submitted Approaches

The AuTexTification shared task received submissions from 36 teams, belonging to 30 different institutions and 18 different countries. All teams participated in the English track of Subtask 1, with 23 teams also taking part in the Spanish track. For Subtask 2, 19 teams participated in the English track and 14 in the Spanish track. Teams were allowed to submit a maximum of 3 runs per subtask and language. Overall, AuTexTification received a total of 175 runs, comprising 71 for the English track of Subtask 1, 47 for the Spanish track, 33 for the English track of Subtask 2, and 24 for the Spanish track. Outside of the competition scope, the AuTexTification datasets have been used in NLP courses within academic institutions. We are aware of at least 3 institutions,[11] with 17 participating teams and 58 runs.

Following the trend in the Natural Language Processing (NLP) field, most teams relied on pre-trained Transformer (Vaswani et al., 2017) models. The most used ones were BERT-based models (Devlin et al., 2019) like *RoBERTa* (Liu et al., 2019) and *DeBERTa* (He, Gao, and Chen, 2021). Also, domain-specific and multilingual variants of *BERT* were frequent, including *XLM-RoBERTa* (Conneau et al., 2020), *RemBERT* (Chung et al., 2020), and *TwhinBERT* (Zhang et al., 2022b). A smaller set of participants included generative models in their systems such as *GPT-2* (Radford et al., 2019), *Grover* (Zellers et al., 2019b), and *OPT* (Zhang et al., 2022a).

Most of the best performing approaches used ensembles of pre-trained models, as well as combinations of lexical, stylometric or statistical features. In some cases, participants fine-tuned their models using auto-train procedures and performed hyper-parameter tuning. Some teams also included *Convolutional Neural Networks* (LeCun et al., 1989) or *Long Short Term Memory (LSTM) Networks* (Hochreiter and Schmidhuber, 1997)

[11] Universitat Politècnica de València, Aix-Marseille Université, and IMT Atlantique.

as part of their systems. Traditional machine learning models like *Logistic Regression* and *Support Vector Machines (SVM)* (Cortes and Vapnik, 1995) were also frequent among the participants. However, these approaches generally performed worse than Transformer-based approaches.

There was also a great diversity in terms of features. Probabilistic token-level features from generative language models seem to play an important role in the best performing approaches. Most participants used contextual representations from pre-trained models, either as features, or through end-to-end fine-tuning. Linguistic features including lexical, structural, and discourse features were also frequent. Among the most common linguistic features, we observed bag of word/char n-grams, counts of personal pronouns, stop-words, punctuations, and POS tags. Some participants also incorporated linguistic and factual knowledge directly in their models. Among these, we found the inclusion of syntactic dependencies in pre-trained models through contrastive learning, Wikipedia fact-checking, and native language identification.

The best ranked systems for each subtask ranged from complex ensembles of many different models and features, to single generative models fine-tuned for the task. In Subtask 1, both for English and Spanish, the best system was proposed by *TALN-UPF* (Przybyla, Duran-Silva, and Egea-Gómez, 2023). This system relied on a bidirectional *LSTM* (Schuster and Paliwal, 1997) model trained with a combination of probabilistic token-level features from different *GPT-2* versions, linguistic token-level features such as word-frequencies or grammar errors, and text representations from pre-trained encoders. Besides, *TALN-UPF* was the only team that considered a cross-domain evaluation in the validation step, by performing cross-validation over topically-split data after inferring the topics using *Latent Dirichlet Allocation* (Blei, Ng, and Jordan, 2003). In the Spanish track, the *TALN-UPF* system performed similar to the *Lingüística_UCM* system (Alonso et al., 2023), consisting of an *SVM* trained with a set of morphological, lexical, and discourse features selected according linguistic expertise and human analysis.

In Subtask 2, both for English and Spanish, the three runs of the *Drocks* team (Abburi et al., 2023) were the highest ranked ones. These systems were ensembles of five different Transformer-based classifiers fine-tuned on the task. The best ensembles differed for each language. For English, the best ensemble was an *Error-Correcting Output Codes* (Dietterich and Bakiri, 1994) model trained using the concatenation of the classification probabilities as features. For Spanish, the best ensemble was implemented with an *SVM* using the average of the classification probabilities as features.

4.2 Baselines

We consider several baselines for each subtask and language. Namely, we include a random baseline (*Random*), zero-shot (*SB-ZS*) and few-shot (*SB-FS*) approaches based on text and label embedding similarities, a bag-of-words encoding with logistic regression (*BOW+LR*), Low Dimensional Semantic Embeddings (*LDSE*), and fine-tuned language specific transformers (Transformer), *DeBERTaV3* (He, Gao, and Chen, 2021)[12] for English and *RoBERTa-BNE* (Fandiño et al., 2022)[13] for Spanish. These baselines consist in the following:

Random. The random baseline assuming class balance. Defined as $\frac{1}{C}$ where C is the number of classes.

SB-ZS and SB-FS. Zero-shot and Few-Shot Symanto Brain API,[14] a ©Symanto solution optimized for highly efficient and scalable state-of-the-art zero-shot and few-shot classification (Mueller, Pérez-Torró, and Franco-Salvador, 2022). We verbalize labels for Subtask 1,[15] but not for Subtask 2 given the anonymity of the classes. For *SB-FS* we use 1024 shots.

BOW+LR. We encode the texts with bag of n-grams, using the top 5K word n-grams, $n \in \{1, 2\}$ and character n-grams, $n \in \{2, \ldots, 6\}$ following (Pizarro, 2019). We train a *Logistic Regression* model offered by scikit-learn (Pedregosa et al., 2011) with default parameters on z-score normalized and concatenated features.

[12] `https://tinyurl.com/debertav3`

[13] `https://tinyurl.com/robertabne`

[14] `https://www.symanto.com/nlp-tools/symanto-brain/`

[15] HUM: "*This text has been written by a human.*" GEN: "*This text has been automatically generated by a bot.*"

Rank	Team	Run	Macro-F_1
1	TALN-UPF	HB_plus	**80.91**
2	TALN-UPF	HB	74.16
3	CIC-IPN-CsCog	run2	74.13
22	turquoise_titans	run1	65.79
23	BOW+LR	baseline	65.78
33	turing_testers	run3	60.64
34	LDSE	baseline	60.35
37	OD-21	run3	59.49
38	SB-FS	baseline	59.44
51	swissnlp_team	run2	57.20
52	Transformer	baseline	57.10
69	UMZ	run1	50.18
70	Random	baseline	50.00
74	SB-ZS	baseline	43.47
77	UAEMex	run1	33.87

Table 4: Ranking of Subtask 1 (English).

Rank	Team	Run	Macro-F_1
1	TALN-UPF	HB_plus	**70.77**
2	Ling_UCM	run1	70.60
3	Transformer	baseline	68.52
20	GLPSI	run3	63.90
21	LDSE	baseline	63.58
25	turing_testers	run1	62.77
26	BOW+LR	baseline	62.40
39	bucharest	run2	56.49
40	SB-FS	baseline	56.05
46	ANLP	run1	51.38
47	Random	baseline	50.00
50	UAEMex	run3	35.17
51	SB-ZS	baseline	34.58
53	LKE_BUAP	run3	31.60

Table 5: Ranking of Subtask 1 (Spanish).

LDSE. We represent texts on the basis of the probability distribution of occurrence of their tokens in the different classes with *LDSE* (Rangel, Franco-Salvador, and Rosso, 2018). We train an *SVM* classifier provided by scikit-learn (Pedregosa et al., 2011) with default parameters.

Transformer. We use the HuggingFace ecosystem (Wolf et al., 2020) to fine-tune a pre-trained Transformer with a randomly initialized classification head for 5 epochs and default hyperparameters. We use a batch size of 32 texts for *DeBERTaV3* and a batch size of 64 for *RoBERTa-BNE*.

4.3 Evaluation

The submissions for both subtasks are evaluated with the Macro-F_1 score. Statistical significance is computed through bootstrapping with replacement at a confidence level of $\alpha = 0.95$ with 1,000 resamples.

4.4 Subtask 1: MGT Detection

For the MGT detection subtask, we received 71 submissions from 36 different teams in English, and 47 submissions from 23 teams in Spanish. Tables 4 and 5 show the top-3 performing teams, the weakest team, as well as the first team that beats each baseline, both for English and Spanish.

The best system was proposed by the *TALN-UPF* team, with 80.91 and 70.77 Macro-F_1 scores in English and Spanish. In English, the best team is significantly better than the second-best ranked team. However, in Spanish there are no significant differences between the two best teams and the best baseline. In Figure 3, we illustrate the rank-ordered Macro-F_1 scores for all the teams in both languages.

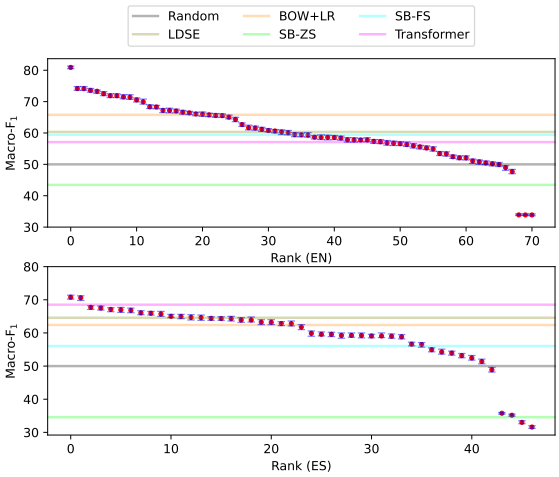
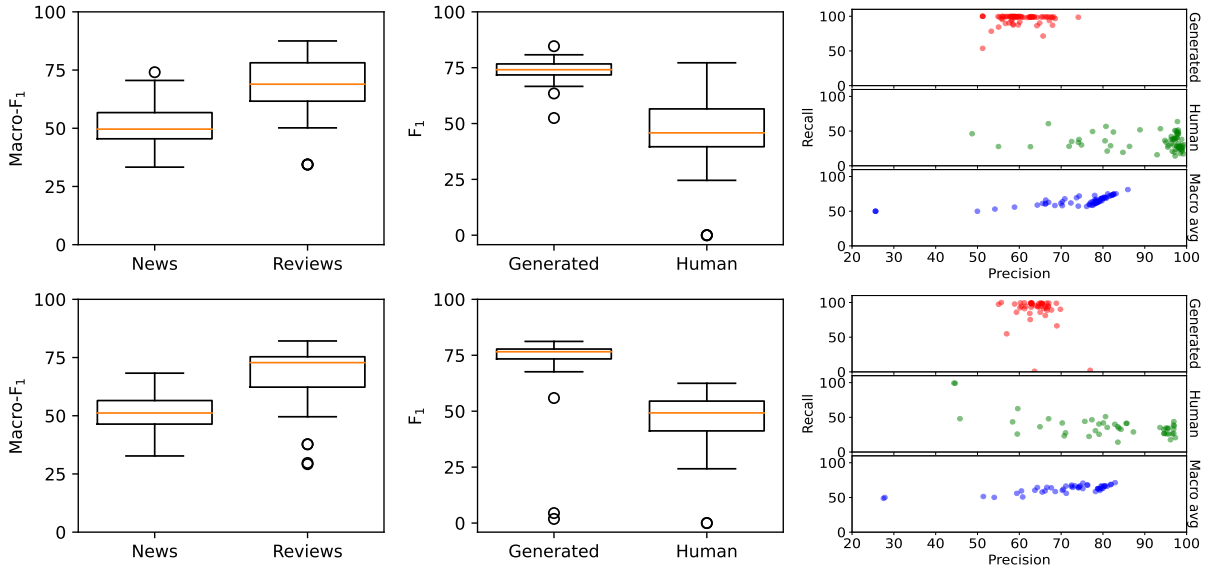


Figure 3: Rank-ordered Macro-F_1 with error bars for Subtask 1 in English (top) and Spanish (bottom). Colored lines are baselines.

Many teams surpassed the best baseline in English by large margins, whereas for Spanish only two teams were able to outperform it with small differences in Macro-F_1. Moreover, the performance of the top-11 ranked teams in English is higher than the performance of the best team in Spanish. This could suggest that detecting MGT and generalizing to new domains is easier in English than in Spanish, either due to language idiosyncrasies or because of the larger availability and quality of English NLP models. For both languages, we observe a linear relationship between the rank-ordered Macro-F_1 scores, with a small set of outliers in both tails. This hints that, even though the resulting Macro-F_1 scores in each language are in different ranges, there is similar variability and difficulty in both languages. The teams' systems cover almost the entire Macro-F_1 range in both languages, and,



(a) Per domain Macro-F_1. (b) Per class F_1. (c) Per class precision-recall plots.

Figure 4: Fine-grained plots for Subtask 1 in English (top) and Spanish (bottom).

in many cases, they are very similar (same Transformer-based models, similar linguistic features, etc.). Therefore, one has to be careful when developing a MGT detector, small changes could lead to large improvements or declines.

We also include fine-grained results per-domain and per-class in Figure 4. When observing the domain-wise Macro-F_1 scores in Figure 4a, we find that the systems generalized better to reviews than to news, with a mean Macro-F_1 below the random baseline for the latter. Furthermore, both domains show long-tailed distributions, revealing the variability in generalization capabilities of the systems. Concerning class-wise F_1 scores in Figure 4b, we find that the systems are better at classifying generated text, and there is lower dispersion among the systems' F_1 scores for this class than for the human class. From the precision-recall distributions depicted in Figure 4c, we observe that systems are more biased towards predicting text to be generated (high recall), often doing so incorrectly (low precision). We observe the opposite for human texts, few predictions (low recall) that are mostly correct (high precision). All the conclusions above hold for both languages.

For the sake of completeness, we refer the reader to the AuTexTification repository,[16] which includes additional plots, the most difficult and easiest examples for the systems, complete rankings including submissions outside the competition, etc.

4.5 Subtask 2: MGT Attribution

For the MGT Attribution subtask we received 33 submissions from 19 different teams in English, and 24 submissions from 14 teams in Spanish. Tables 6 and 7 show the top-3 performing teams, the weakest team, as well as the first team that beats each baseline, both for English and Spanish.

The best system was submitted by team *Drocks*, obtaining 62.5 and 65.37 Macro-F_1 scores for English and Spanish, respectively. This is in contrast to the best scores of Subtask 1 nearing 80 and 70 Macro-F_1, showing that in-domain MGT attribution is more difficult than out-of-domain MGT detection. In this subtask, teams did not deviate significantly from the baselines, and for both languages the relative ranking of baselines remained the same, as opposed to Subtask 1. Rank ordered Macro-F_1 scores for both languages are presented in Figure 5. Few teams were able to surpass the best baselines, with most submissions performing between the top-2 baseline scores. Similarly to Subtask 1, we observe a linear relationship between rank and Macro-F_1 with outliers in the right tail, meaning that there is variability and difficulty in attributing MGT irrespective of language. However, teams cover

[16] https://tinyurl.com/overview-results

Rank	Team	Run	Macro-F_1
1	Drocks	run3	**62.50**
2	Drocks	run1	61.29
3	Drocks	run2	61.27
4	ViDa	run1	60.99
5	Transformer	baseline	60.42
31	LKE_BUAP	run1	45.62
32	LDSE	baseline	44.56
33	turquoise_titans	run2	43.37
34	BOW+LR	baseline	39.98
35	UAEMex	run2	33.19
36	SB-FS	baseline	28.94
37	Random	baseline	16.66
38	SB-ZS	baseline	15.70
39	ANLP	run1	14.61

Table 6: Ranking Subtask 2 (English).

Rank	Team	Run	Macro-F_1
1	Drocks	run2	**65.37**
2	Drocks	run3	64.72
3	Drocks	run1	64.17
7	TALN-UPF	Hybrid_plus	61.45
8	Transformer	baseline	61.34
20	iimasPLN	run1	51.43
21	LDSE	baseline	45.46
22	BOW+LR	baseline	45.31
25	UAEMex	run2	33.78
26	SB-FS	baseline	31.38
28	ANLP	run1	17.93
29	Random	baseline	16.66
30	SB-ZS	baseline	16.23

Table 7: Ranking Subtask 2 (Spanish).

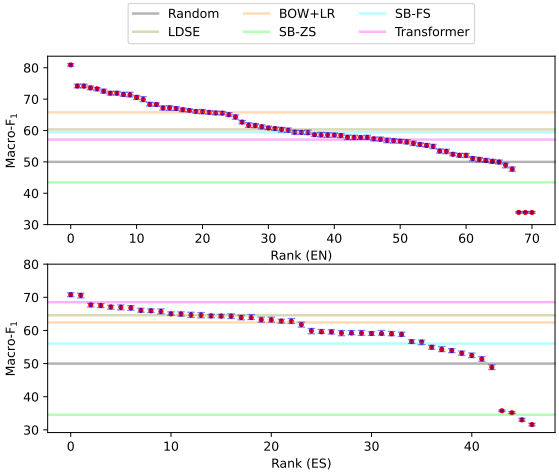


Figure 5: Rank-ordered Macro-F_1 for Subtask 2 in English (top) and Spanish (bottom). Colored lines are baselines.

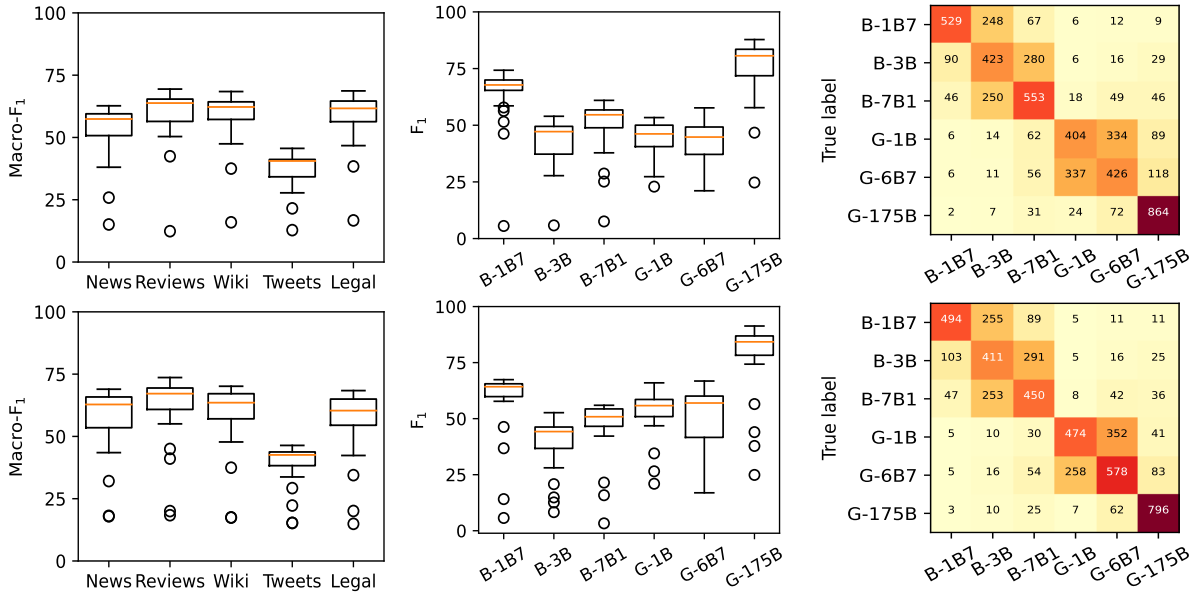
a smaller range of Macro-F_1 scores than in Subtask 1, suggesting there is less variability when attributing MGT than detecting it. In contrast to Subtask 1, teams generally obtained better Macro-F_1 scores in Spanish than English, but the differences were marginal, which could be because of the randomness of the learning procedures or due to a smaller number of participants in Subtask 2. Generally, MGT attribution appears promising but limited, suggesting the need for further research into new approaches or framings of the problem. Fine-grained per-domain and per-class results for Subtask 2 are presented in Figure 6. Per-domain results (Figure 6a) show that attribution of generated tweets is much more difficult than the remaining domains. For tweets, systems are unable to reach 50% Macro-F_1, while for the other domains they surpass it by a large margin. We additionally find many outliers toward lower scores, indicating the difficulty of the task. Finally, most domains have similar distributions centered around different medians, meaning that the variability of participating systems is maintained through all five domains. We also present per-class results in Figure 6b, where we find that it is easier to attribute generated text to *BLOOM-1B7* and *text-davinici-003*. Moreover we observe large variability for *curie*, while the other classes have narrower distributions.

Additionally, we computed overall confusion matrices by taking the median at each position of the confusion matrix from all the participant's systems. Figure 6c shows the results for English and Spanish. In both languages, the largest confusions are across models within the same families, suggesting that it is easier to distinguish generation models of different families. Besides, *text-davinci-003* is the model with less number of confusions, being different enough to be easily distinguished from the other models.

Once again, we refer to the AuTexTification repository[16] for additional plots, results and analyses.

5 *Conclusions and Future Work*

This paper describes the AuTexTification shared task at IberLEF 2023, which aimed to study the automatic detection of MGT in cross-domain scenarios and MGT attribution to specific generation models, across five domains and two languages. The AuTexTification dataset was comprised of around 160,000 texts collected through an automatic data gathering process which can be easily extended to new domains and languages. The task received a significant amount of participation: 175 runs from 36 teams, belonging to 30 different institutions and 18 different



(a) Per domain Macro-F_1. (b) Per class F_1. (c) Median confusion matrices.

Figure 6: Fine-grained plots for Subtask 2 in English (top) and Spanish (bottom). B- prefix denotes *BLOOM* models and G- prefix denotes *GPT* models.

countries, thus showing the overall interest of the community in addressing MGT detection and attribution. Moreover, other 17 teams submitted 58 runs although after the deadline, for a total of 233 runs by 53 teams.

The participating systems relied on a wide variety of approaches, with a strong trend towards the use of pre-trained Transformer models. Ensembles of pre-trained models and combinations of probabilistic, lexical, and stylometric features led to the best performing systems in both subtasks. The results suggest that cross-domain MGT detection is easier in English than in Spanish, and that MGT attribution is generally more challenging than MGT detection. While MGT attribution appears promising, the small gap between the participant's systems and the baselines encourage further research. Overall, the results suggest that MGT detection and attribution remain challenging tasks and there is potential for further progress.

As future work, we hope to expand the AuTexTification dataset to include more languages, domains, generation models and decoding strategies, to encourage the development of more robust and generalizable systems. Furthermore, it would be valuable to explore alternative formulations of MGT attribution, as fine-grained attribution remains a challenging task.

Acknowledgements

We would like to thank Guillermo Pérez-Torró, Ian Borrego Obrador, and Angelo Basile for their precious help participating in the human assessment, and Mara Chinea Rios for developing a custom implementation of the LDSE baseline.

The work from Symanto has been partially funded by the Pro^{2}Haters - Proactive Profiling of Hate Speech Spreaders (CDTi IDI-20210776), the XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics (MICIN PLEC2021-007681), the OBULEX - *OBservatorio del Uso de Lenguage sEXista en la red* (IVACE IMINOD/2022/106), and the ANDHI - ANomalous Diffusion of Harmful Information (CPP2021-008994) R&D grants. The work of Areg Mikael Sarvazyan has been partially developed with the support of valgrAI - Valencian Graduate School and Research Network of Artificial Intelligence and the Generalitat Valenciana, and co-founded by the European Union. The research at the Universitat Politècnica de València was framed under the FairTransNLP research project, Grant PID2021-124361OB-C31 funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe.

References

Abburi, H., M. Suesserman, N. Pudota, B. Veeramani, E. Bowen, and S. Bhattacharya. 2023. Generative ai text classification using ensemble llm approaches. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023). CEUR Workshop Proceedings, CEUR-WS, Jaén, Spain*.

Alonso, L., J. A. Gonzalo, A. M. Fernández-Pampillón, M. Fernández, and M. V. Escadell-Vidal. 2023. Using linguistic knowledge for automated text identification. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023). CEUR Workshop Proceedings, CEUR-WS, Jaén, Spain*.

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*.

Barbieri, F., L. Espinosa Anke, and J. Camacho-Collados. 2022. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Language Resources and Evaluation Conference*, pages 258–266, Marseille, France, June. European Language Resources Association.

Benjamin, M., P. Gagnon, N. Rostamzadeh, C. Pal, Y. Bengio, and A. Shee. 2019. Towards standardization of data licenses: The montreal data license. *arXiv preprint arXiv:1903.12262*.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chalkidis, I., M. Fergadiotis, and I. Androutsopoulos. 2021. Multieurlex – a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Chung, H. W., T. Fevry, H. Tsai, M. Johnson, and S. Ruder. 2020. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

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

Contractor, D., D. McDuff, J. K. Haines, J. Lee, C. Hines, B. Hecht, N. Vincent, and H. Li. 2022. Behavioral use licensing for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 778–788.

Cortes, C. and V. Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dietterich, T. G. and G. Bakiri. 1994. Solving multiclass learning problems via error-correcting output codes. *Journal of artificial intelligence research*, 2:263–286.

Eloundou, T., S. Manning, P. Mishkin, and D. Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Fandiño, A. G., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

González, M. D., E. Martínez-Cámara, M. Martín-Valdivia, and L. López. 2014. Cross-domain sentiment analysis using

spanish opinionated words. volume 8455, pages 214–219, 06.

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. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Hochreiter, S. and J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Holtzman, A., J. Buys, L. Du, M. Forbes, and Y. Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

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

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

Kashnitsky, Y., D. Herrmannova, A. de Waard, G. Tsatsaronis, C. C. Fennell, and C. Labbé. 2022. Overview of the dagpap22 shared task on detecting automatically generated scientific papers. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 210–213.

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.

LeCun, Y., B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. 1989. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2.

Leis, A., F. Ronzano, M. A. Mayer, L. I. Furlong, and F. Sanz. 2019. Detecting signs of depression in tweets in spanish: Behavioral and linguistic analysis. *J Med Internet Res*, 21(6):e14199, Jun.

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*.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Maronikolakis, A., H. Schütze, and M. Stevenson. 2021. Identifying automatically generated headlines using transformers. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 1–6, Online, June. Association for Computational Linguistics.

McAuley, J. and J. Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, page 165–172, New York, NY, USA. Association for Computing Machinery.

Mitchell, E., Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Mueller, T., G. Pérez-Torró, and M. Franco-Salvador. 2022. Few-shot learning with siamese networks and label tuning. In *Proceedings of the 60th Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8532–8545.

Naji, I. 2012. TSATC: Twitter Sentiment Analysis Training Corpus. In *thinknook*.

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. *ArXiv*, abs/1808.08745.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Gray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. 2022. Training language models to follow instructions with human feedback. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*.

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.

Pizarro, J. 2019. Using n-grams to detect bots on twitter. In *Conference and Labs of the Evaluation Forum*.

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*.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Rangel, F., M. Franco-Salvador, and P. Rosso. 2018. A low dimensionality representation for language variety identification. In *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part II 17*, pages 156–169. Springer.

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

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

Schuster, M. and K. K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Scialom, T., P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online, November. Association for Computational Linguistics.

Seger, E., A. Ovadya, B. Garfinkel, D. Siddarth, and A. Dafoe. 2023. Democratising ai: Multiple meanings, goals, and methods. *arXiv preprint arXiv:2303.12642*.

Shamardina, T., V. Mikhailov, D. Chernianskii, A. Fenogenova, M. Saidov, A. Valeeva, T. Shavrina, I. Smurov, E. Tutubalina, and E. Artemova. 2022. Findings of the the ruatd shared task 2022 on artificial text detection in russian. *arXiv preprint arXiv:2206.01583*.

Spitale, G., N. Biller-Andorno, and F. Germani. 2023. Ai model gpt-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Uchendu, A., T. Le, K. Shu, and D. Lee. 2020. Authorship attribution for neural text generation. In *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.

Widder, D. G., D. Nafus, L. Dabbish, and J. Herbsleb. 2022. Limits and possibilities for "ethical ai" in open source: A study of deepfakes. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2035–2046.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. 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.

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

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

Zhang, S., S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. 2022a. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, X., Y. Malkov, O. Florez, S. Park, B. McWilliams, J. Han, and A. El-Kishky. 2022b. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*.