

symanto

Intro to Symanto Research



SYMANTO AT A GLANCE

We are a software- and service provider around **text analytics**, combining **leading edge technology** with many years of **insights expertise** since 2010.

+70 employees

thereof, ...

+29

- **Psychologists**
- **Data Scientists**
- **Research Scientists**
- **Research Engineers**

and

- **Marketing- & Market Research Industry Experts**

from 32 nationalities

+90 research papers &
10 PhDs

Germany (HQ)
Spain



Collaboration with professors,
researchers in Europe & US

Germany



Spain



Switzerland



UK



US



Partnerships with



Leading biomedical
research institution of
the German federal
government



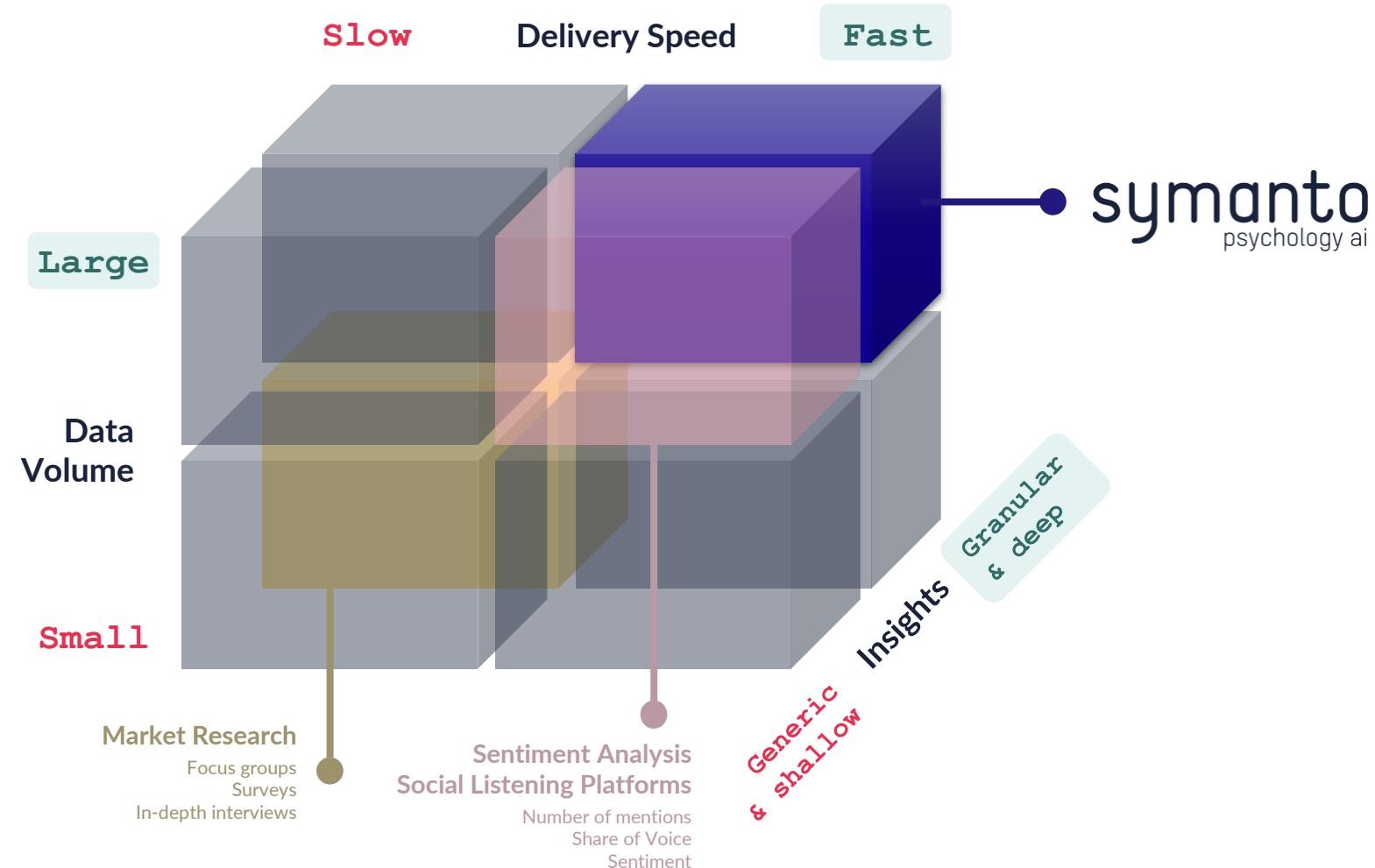
Non-profit association
that focusses on
democratizing AI
throughout Spain

Business partners such as ...



SWEET SPOT BETWEEN TECH & INSIGHTS

Fulfilling the needs of faster & deeper insights through AI based tech



DELIVERY

Insights are delivered by a scalable platform, as one-off projects, on-going monitoring or as a SaaS or AlaaS solution.

QUESTIONS / HYPOTHESES

Outside In client & peers	Public & Proprietary online sources Gain access to > 1.500 data sources <ul style="list-style-type: none"> • Social media • Review sites • Forums • News • Employee portals • Company DBs • Clickstream data • Location data • Search data • Pricing data
Inside In client	Internal data sources Consolidate all text and image data sources <ul style="list-style-type: none"> Survey open-ends NPS Call centre Chatbots

DATA SOURCES

PROPRIETARY AI MODELS

Sentiment & Psychographic analysis available
in more than 50 languages

Bot Detection	Emotion	Custom Model Creation
Spam Detection	Age Gender	Mental Health
Topic Extraction	Personality Traits	Dataless Image & Text Classification
Aspect-based Sentiment	Communication Style	... and more

■ Unique models only provided by Symanto

DELIVERY

- One-off projects**
to answer your questions
- Ongoing monitoring**
to continuously improve
- SaaS or AlaaS**
to enhance your own analytics team

SELECTED GLOBAL CLIENTS

Automotive	 VOLKSWAGEN FINANCIAL SERVICES AKTIENGESELLSCHAFT		al tayer motors	الطالب للسيارات
Consulting & Private Equity	Delivered more than 300 projects since 2021 to Top Management Consulting & Private Equity firms globally			
Consumer	 harman/kardon® by HARMAN			 HUBLOT NEWYORKER
FMCG	 Nestlé			
Pharma	 medac			
Media & Entertainment	 M&CSAATCHI			 RTL ZWEI
Retail	 Boots			
Sports & Gaming	 FC BAYERN MÜNCHEN	 ACTIVISION		reworks XBOX
Telco & Tech	 Türk Telekom			
Other B2C	 NÜRNBERGER VERSICHERUNG	HOFBRÄU MÜNCHEN WIRTSHAU	Stadtverkehr für Hildesheim	
B2B	 KWS			

SELECTED CASE STUDIES

Brand Health

Identification of functional & emotional brand equities from consumers' POV and measurement of Brand Love and its drivers

JBL (Consumer Electronics)



Customer Experience

CX analytics led product development
Harman (Consumer & B2B Electronics)



ESG Scan

Consumer perception towards Company's ESG, benchmarking against peers & gap analysis (reports vs perception)

Beauty Retailer



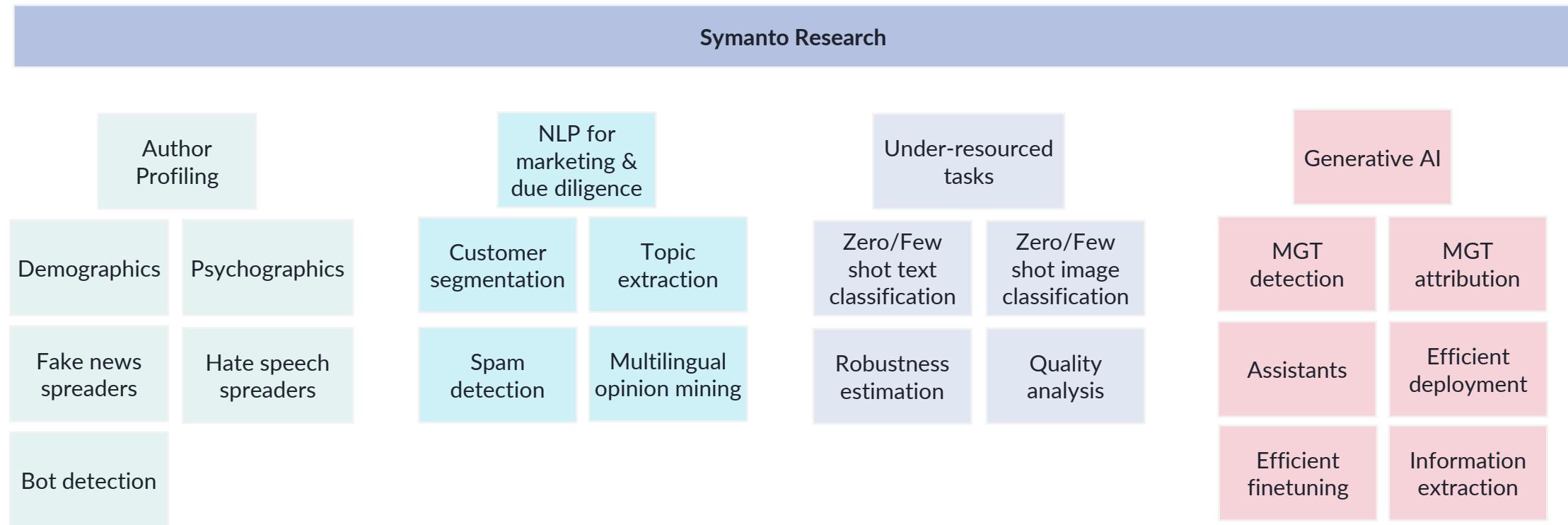
Trend Analysis

Identification of consumer needs driven new trends and innovation areas

Dr. Oetker (FMCG)



MAIN RESEARCH LINES



More than 30 publications just in these lines.

SYMANTO PSYCHOLOGY AI

LLM Refresher

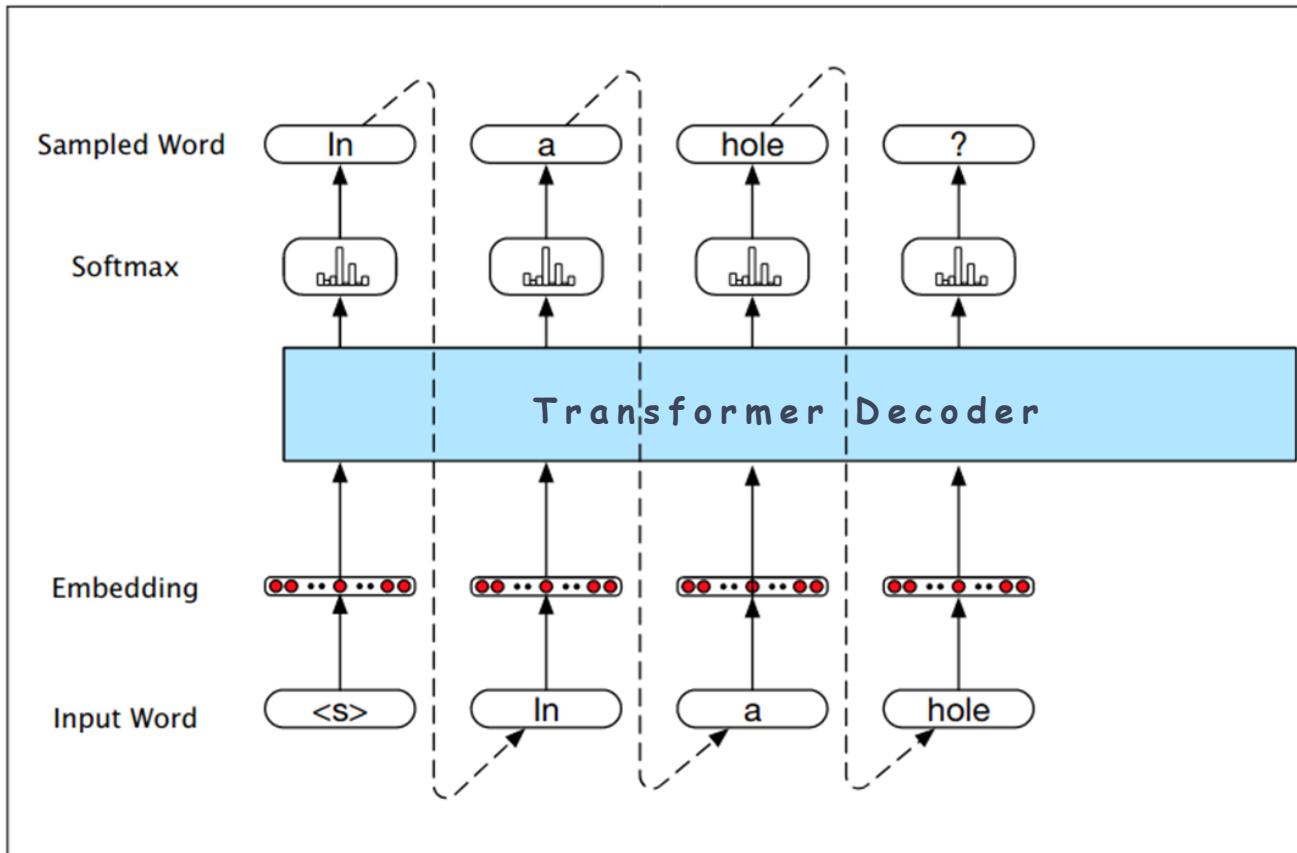
Public



Neural-based Causal Language Models

- Compute the **probability** of a **token sequence** $p(X) = p(x_1, x_2, x_3, \dots, x_T)$
- From the token probability, at each step t , given its **previous context**, $p(x_t|x_1, x_2, x_3, \dots, x_{t-1})$
- By means of the chain rule: $p(X) = p(x_1) \prod_{t=2}^T p(x_t|x_1 \dots x_{t-1})$
- $p(x_t|x_1, x_2, x_3, \dots, x_{t-1})$ is computed using a neural network
- **Autoregressive (unidirectional)** processing, given the prefix until the step t , compute a **probability distribution** over the **vocabulary**
- Not necessary to consider the **Markov assumption** (don't forget the distant past)
- Current state-of-the-art architecture: **Transformer decoder**
- Models: GPT, LLaMA, Mixtral, Claude, ...

Inference process

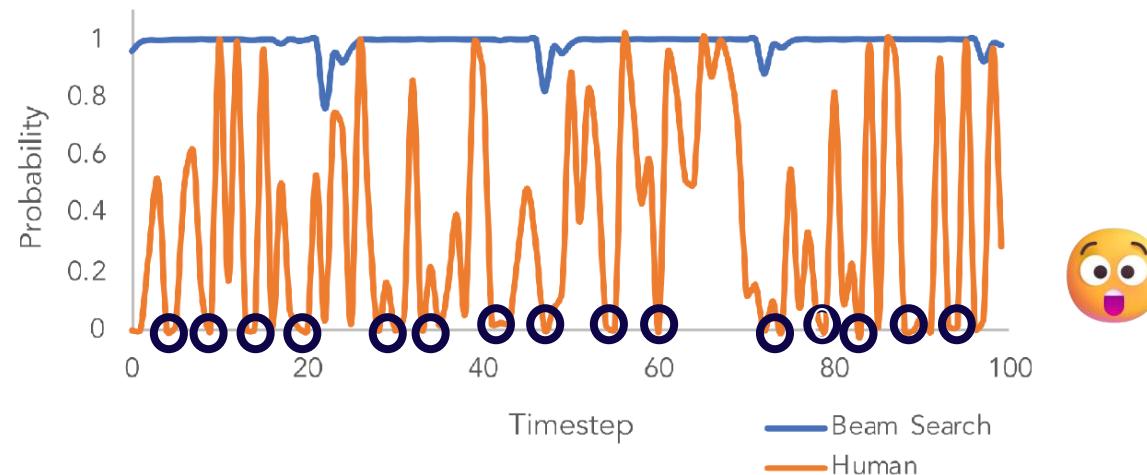


- Once trained, they can **generate text** (**decode**) by means of **sampling** the next token distribution at each timestep t
- We will refer a generated text as **MGT** along this session

Decoding

- We can generate by finding the **most likely suffix** given a prefix (e.g. beam search or greedy):

$$\mathbf{w}_t^{T*} = \operatorname{argmax}_{\mathbf{w}_t^T} p(\mathbf{w}_t^T | \mathbf{w}_1^t)$$

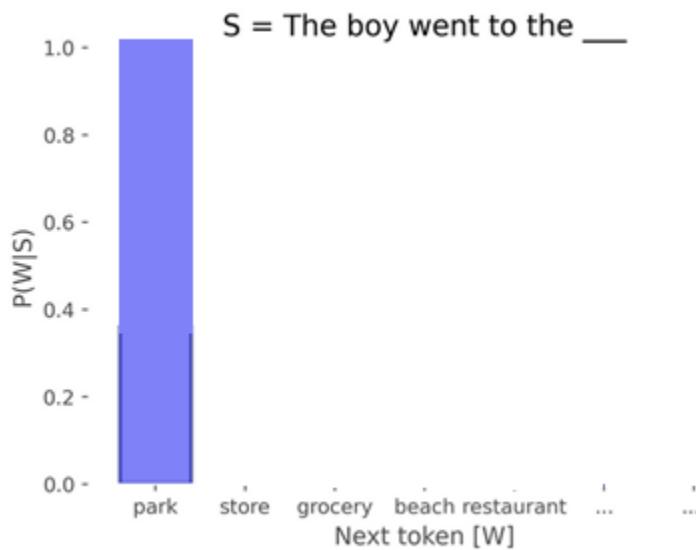


- But humans do not write according to maximum probability, we write **surprising**, **coherent** and **diverse** text
- Decoding by means of **sampling**

Decoding: softmax with temperature

➤ Softmax with temperature:

$$\frac{\exp\left(\frac{h^T w}{\tau}\right)}{\sum_{w_i \in V} \exp\left(\frac{h^T w_i}{\tau}\right)}$$



Temperature=0:

- Peak distribution.
- Determinism.



Temperature=1:

- Distribution directly from LLM.



Temperature=2:

- Near uniform distribution.
- Random¹² generations.

How to detect MGT? Test yourself...

Can you spot the generated text? 🤔

I recently purchased a pair of Celestron Skymaster Binoculars and I am quite impressed with their performance. They are solidly built and provide a good balance of size and weight. The optics are sharp and the image is bright. I was able to easily spot many celestial objects in the night sky, including planets, stars, and galaxies. The focusing mechanism is smooth and easy to use.

Después de la entrevista, Isabel Rodríguez se reunió con los miembros del PSOE para discutir su postura en relación a las enmiendas presentadas al proyecto de ley. La reunión fue productiva y los miembros del partido acordaron apoyar algunas de las enmiendas presentadas por el PP, siempre y cuando se ajustaran a los principios y valores del PSOE. Rodríguez señaló que el partido continuará trabajando en conjunto con el PP en el interés del bienestar del país.

How to detect MGT?

- Machine-aided [1]: capture text artifacts automatically to help humans detect MGT
- Watermarking [2]: make MGT self-identifiable through cryptographic watermarks
 - Only possible if everyone enforces watermarks (otherwise: can paraphrase with another model)
- Zero-shot [3] (white-box)
 - Use LLM probabilities to detect its own MGT: **not generalizable to new models**
 - We usually don't know what models generated the texts
 - Could not have white-box access to it

[1] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *Advances in neural information processing systems*, 32.

[2] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A Watermark for Large Language Models. *International Conference on Machine Learning*.

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

How to detect MGT?

- Supervised [4, 5]
 - Train models on text and its linguistic and statistical features: **generalization is possible**
 - Need high quality multi-domain/style datasets
 - Transformer-based models studied under single-domain assumption
 - Generalization capabilities to new domains must be studied [6]
- We will focus on this one along this session

[4] Rodriguez, J., Hay, T., Gros, D., Shamsi, Z., & Srinivasan, R. (2022, July). Cross-Domain Detection of GPT-2-Generated Technical Text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1213-1233).

[5] Maronikolakis, A., Schütze, H., & Stevenson, M. (2021, June). Identifying Automatically Generated Headlines using Transformers. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda* (pp. 1-6).

[6] Sarvazyan, A. M., González, J., Franco-Salvador, M., Rangel, F., Chulvi, B., & Rosso, P. (2023). Overview of AuTexTification at IberLEF 2023: Detection and Attribution of Machine-Generated Text in Multiple Domains. *Procesamiento Del Lenguaje Natural*, 71, 275-288.

SYMANTO PSYCHOLOGY AI

AuTexTification: Detection and Attribution of Machine-Generated Text in Multiple Domains



Areg Mikael Sarvazyan¹, Jose Ángel González¹, Marc Franco-Salvador¹, Francisco Rangel¹, Berta Chulvi² and Paolo Rosso²

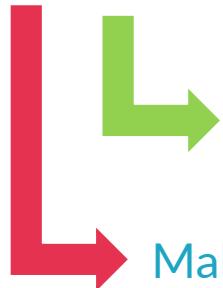
¹ Symanto Research Spain

² Universitat Politècnica de València



Background

- Large scale democratization of highly capable **Large Language Models (LLMs)** [7]



Improves workflows of individuals and organizations

Malicious users can use them to generate text massively (fake news, opinions, propaganda, etc.) [8]

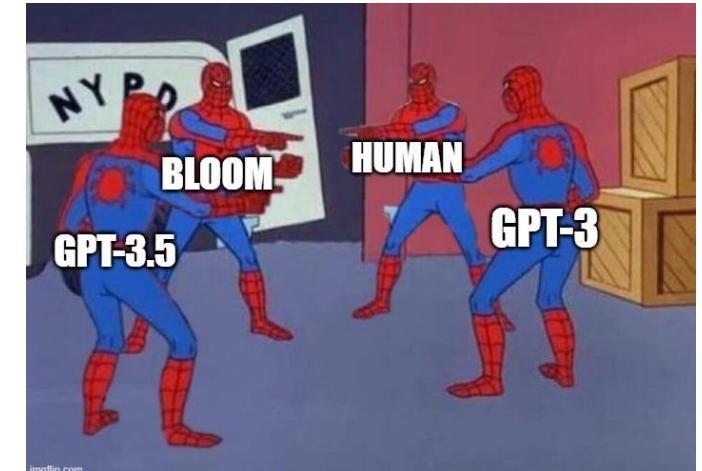
- Entry **barriers** are becoming lower, and generated **text quality** is increasing:

- Need to deal with **malicious applications** of LLMs



Motivation

- Ensure a **responsible usage** of LLMs through:
 - Content moderation of machine-generated text (MGT) with malicious purposes
 - Checking compliance of AI ethics, regulations [9] and licenses [10]
- We aim to boost the development of language technologies for:
 - MGT Detection: distinguishing between **MGT** and **human text**
 - MGT Attribution: attributing a **MGT** to a particular **LLM**



AuTexTification task



AuTexTification focus

- Focuses on a specific kind of **generation campaigns**:
- **Massive** quantities of text
- Without large **computational resources**
- Without **technical knowledge**
- Without **human intervention**

		Pre-trained	Fine-tuned
Accessibility	Not modified by a human	Everyone	Only technical
Computational Resources		Low	High
Human Resources		Low	Low
Generation Scale		High	High
Generation Quality		Medium	High
Accessibility	Modified by a human	Everyone	Only technical
Computational Resources		Low	High
Human Resources		High	High
Generation Scale		Low	Low
Generation Quality		High	Perfect

- **Supervised** detection [11] and attribution [12]: the only approach **generalizable** to new LLMs
- **Multi-domain**: MGT detectors are known to not generalize well across domains [13]
- **Multiple languages**: generation campaigns could be performed in any language

[11] Ippolito, D., D. Duckworth, C. Callison- Burch, and D. Eck. 2020. Automatic detection of generated text is easiest when humans are fooled.

[12] Uchendu, A., T. Le, K. Shu, and D. Lee. 2020. Authorship attribution for neural text generation.

[13] Jawahar, G., Abdul-Mageed, M., & Lakshmanan, L. (2020). Automatic Detection of Machine Generated Text: A Critical Survey.

AuTexTification subtasks

AuTexTification consisted of **two subtasks**:

Subtask 1: MGT detection

Participants' systems had to determine whether a text had been written by a human or by an LLM, in a **cross-domain** setting (three domains were used for training, and two different domains for testing)

Binary classification task

Macro- F_1

Subtask 2: MGT attribution

Assuming a text has been identified as MGT, participants' systems had to determine what LLM generated that text.

Multi-class classification task (6 classes)

Macro- F_1

AuTexTification datasets



AuTexTification datasets

- The AuTexTification data consists of:
 - Human text, extracted from existing datasets
 - MGT, completions from human prefixes
- Two languages: Spanish and English
- Five domains: tweets, reviews, how-to, news, and legal articles
- Six LLMs:
 - BLOOM (1b1, 3b, 7b1),
 - GPT (Babbage, curie, text-davinci-003)
- Fully balanced (human/MGT, domains, and LLMs)

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

Subtask 1 dataset

Human and MGT 🤖👤
 Train: tweets, how-to, legal
 Test: news, reviews

Subtask 2 dataset

Only MGT 🤖
 Train/test splits (80%/20%)

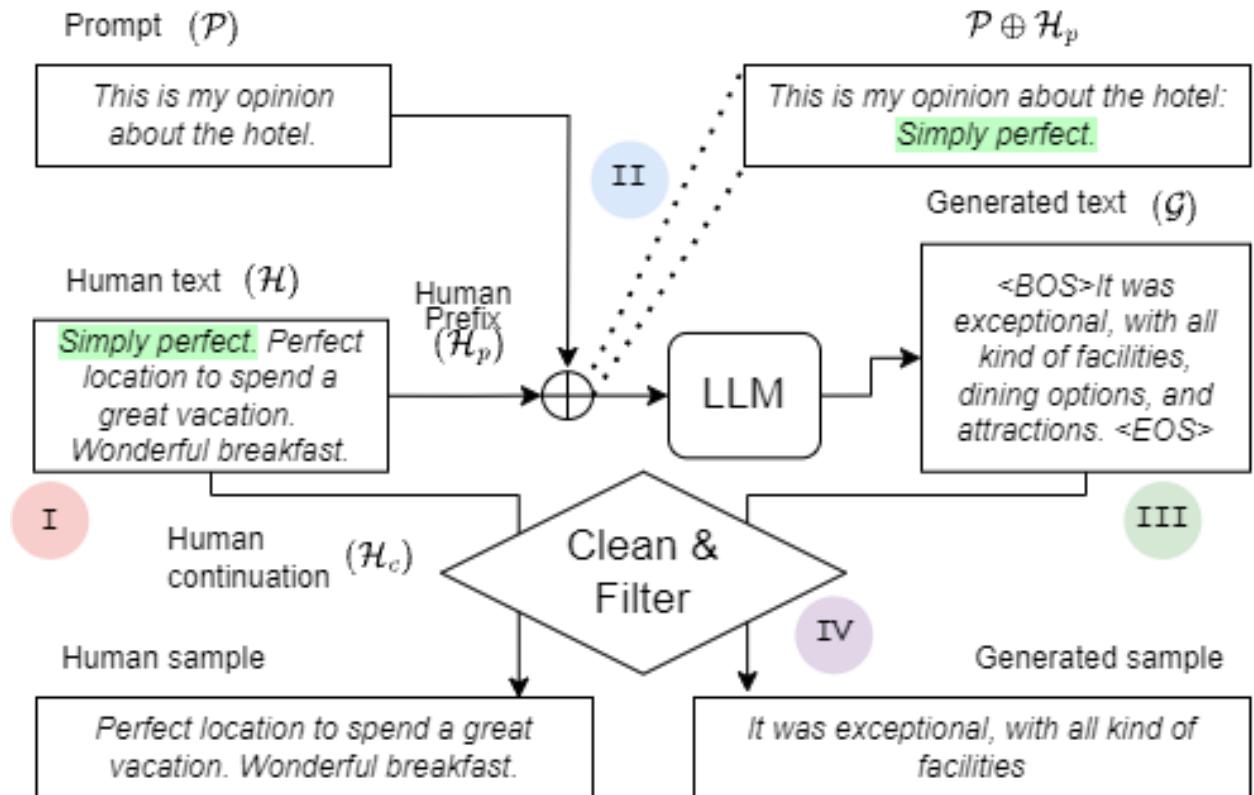


AuTexTification datasets (statistics)

		Subtask 1			Subtask 2					
		GEN	HUM	Σ	BLOOM			GPT		
Spanish	Legal	4,846	4,358	9,204	640	665	712	919	942	919
	News	5,514	5,223	10,737	839	860	881	972	978	987
Spanish	Reviews	5,695	3,697	9,392	952	962	935	945	941	947
	Tweets	5,739	5,634	11,373	967	965	965	928	930	964
	How-to	5,690	5,795	11,485	894	929	960	970	983	966
	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
	News	5,464	5,464	10,928	747	854	906	983	984	984
	Reviews	5,726	5,178	10,904	944	946	939	977	974	972
	Tweets	5,813	5,884	11,697	987	968	980	951	963	969
	How-to	5,862	5,918	11,780	962	976	982	993	993	963
	Total	27,989	27,688	55,677	4,449	4,523	4,639	4,794	4,801	4,815
										28,021

AuTexTification datasets (data gathering)

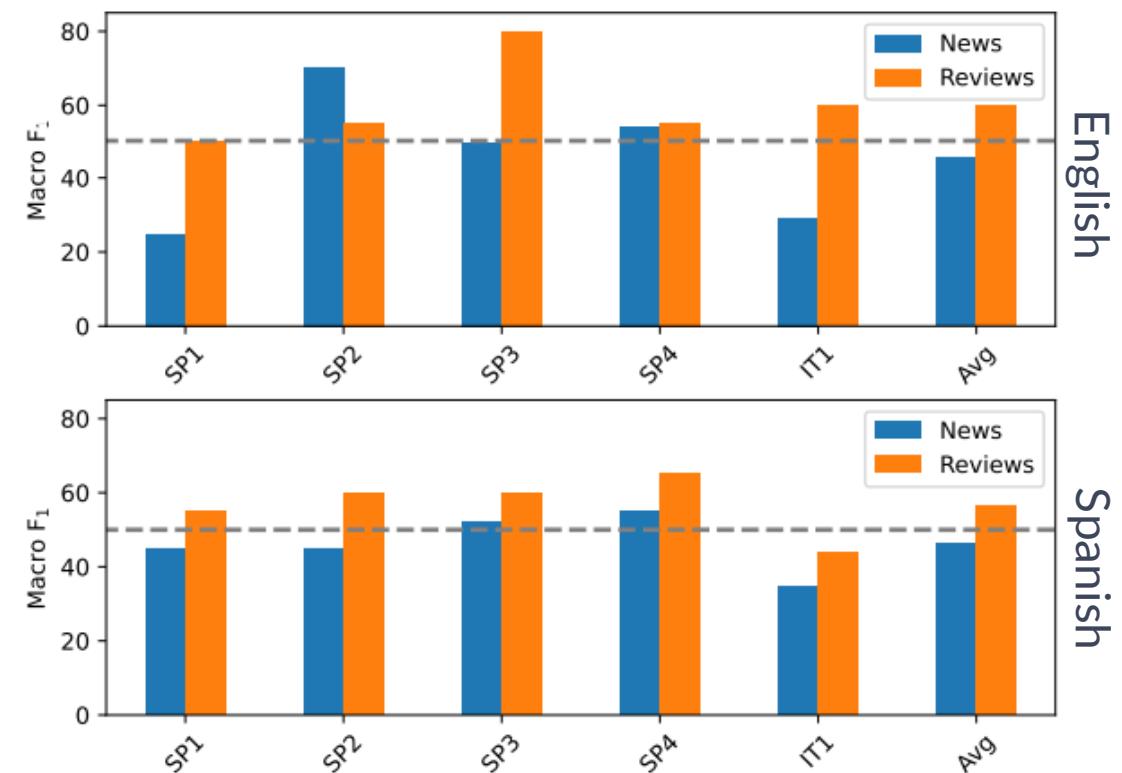
1. Gather a **human text** \mathcal{H} and split into **prefix** \mathcal{H}_p and **continuation** \mathcal{H}_c
2. Concatenate a **prompt** \mathcal{P} to get the **input** to LLMs, $\mathcal{P} \oplus \mathcal{H}_p$
3. Feed to LLM and get **generation** \mathcal{G}
4. **Clean** texts and **filter** by language ID
5. \mathcal{G} is **MGT** and \mathcal{H}_c is **human**



➤ Careful **human inspection** and decoding details to avoid inducing **biases** (length, topics, structure, etc.)

Human assessment of MGT detection

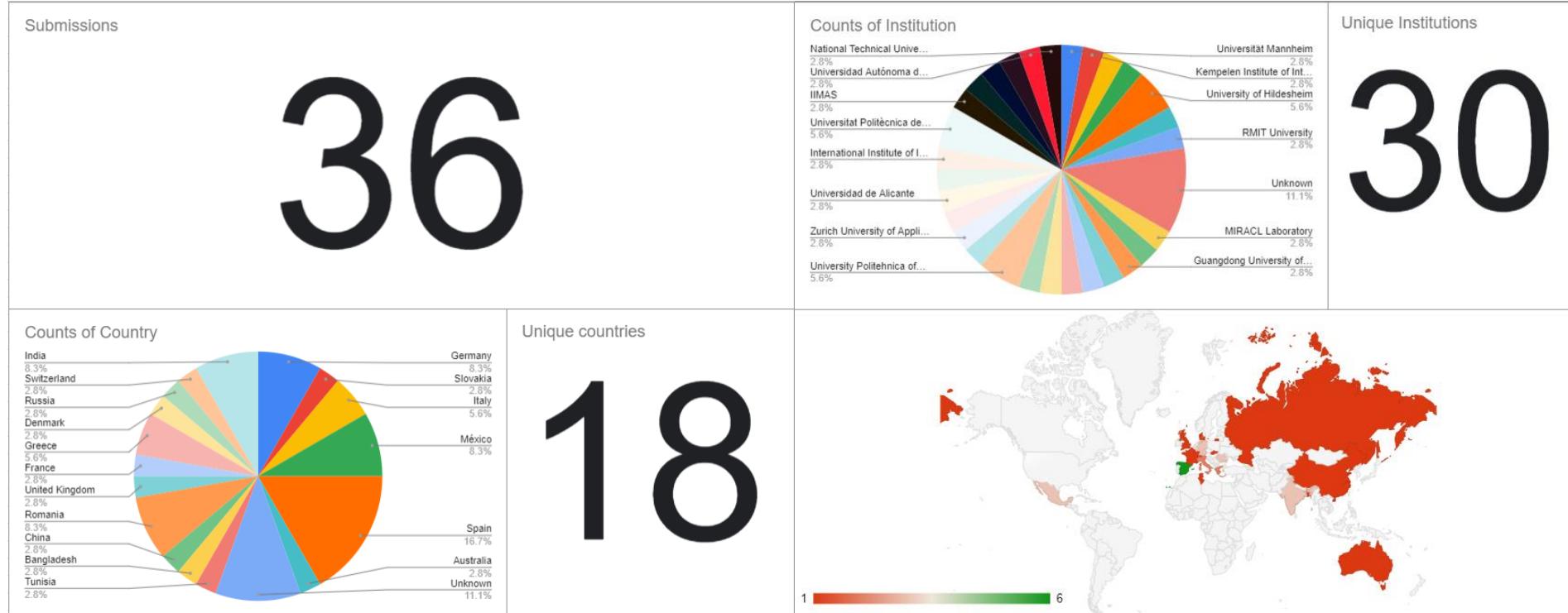
- Small-scale **human study** to assess the difficulty:
 - 5 annotators (4 SP, 1 IT)
 - 40 samples (news/reviews, 20 human, 20 MGT)
- Conclusions:
 - Close to the **random detector**
 - MGT **easier** to detect in reviews than in news
 - **Language proficiency** plays a role
 - Didn't find **any systematic pattern**



Systems in AuTexTification



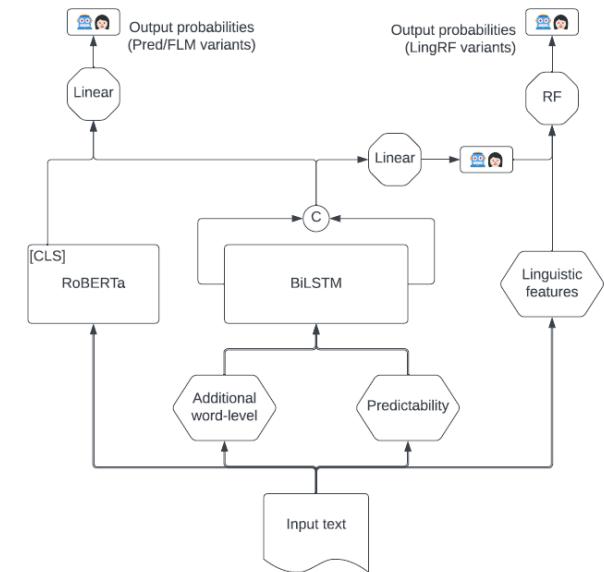
Submitted approaches



- AuTexTification received a total of **175 runs** for all the subtasks and languages

Best approaches

- For Subtask 1 (MGT detection), both for English and Spanish [14]:
 - Bidirectional LSTM trained with:
 - Probabilistic token-level features from different GPT-2 versions
 - Linguistic token-level features
 - Contextual representations from pretrained encoders
- For Subtask 2 (MGT attribution), both for English and Spanish [15]:
 - Ensemble of five different Transformer classifiers finetuned on the task
 - Concatenation of probabilities for English
 - Average of probabilities for Spanish

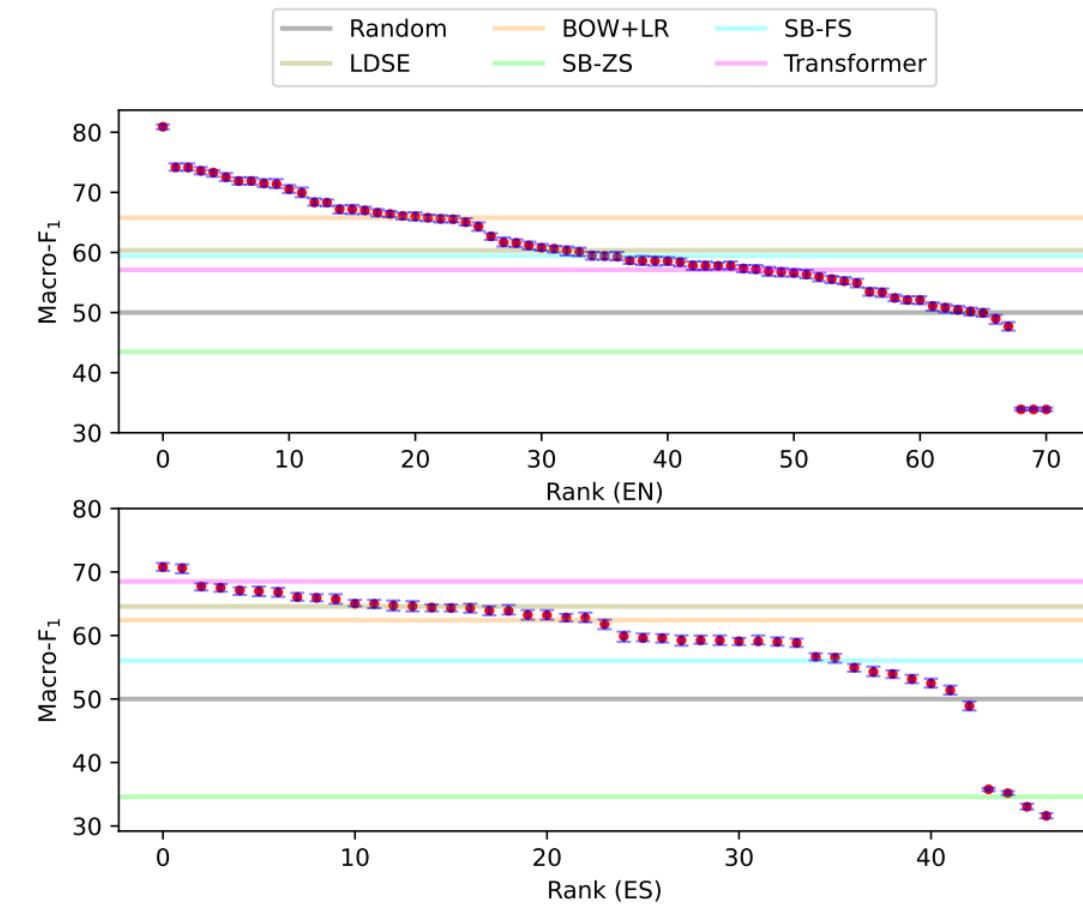


Subtask 1 results



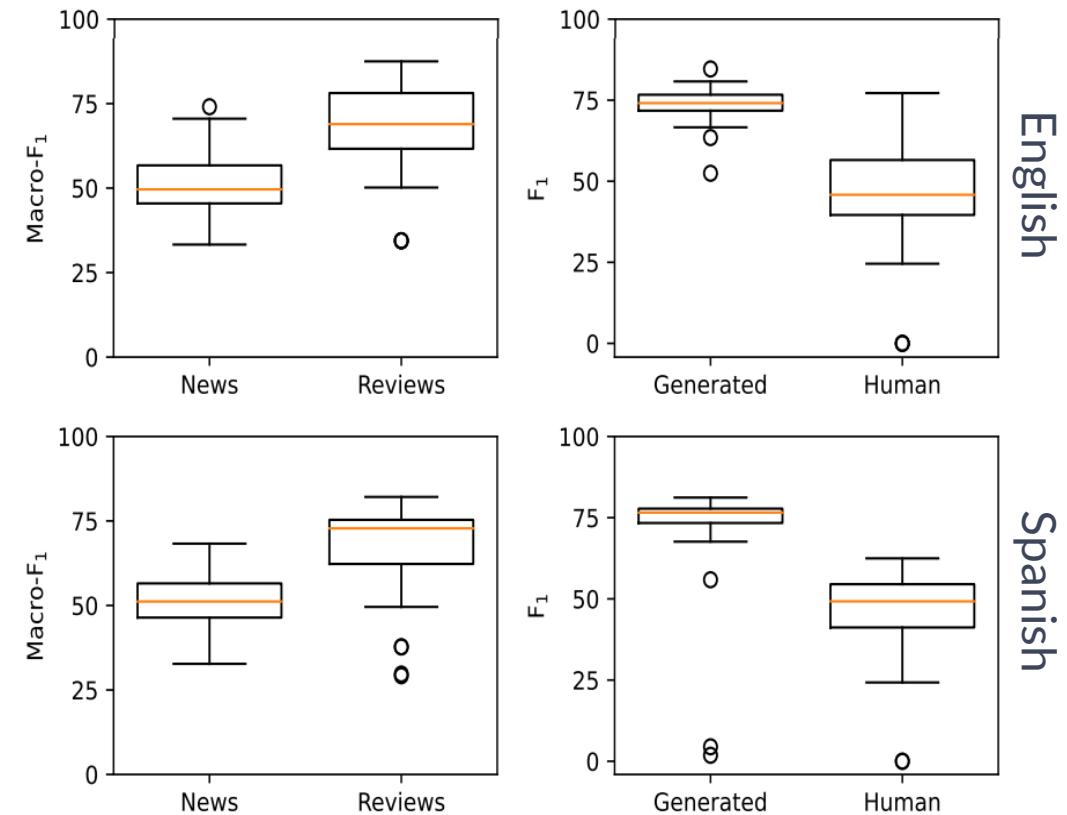
Subtask 1 results (overall)

- Top performance in MGT detection: 70-80 Macro F₁
- Similar variability and difficulty in both languages
- Only two systems are better than the baseline (Spanish)
- Small modeling decisions lead to huge improvements
- Significant gap between 1st and 2nd in English (same system but with additional token-level features)



Subtask 1 results (fine-grained)

- Systems **generalize** better to reviews than to news:
 - Almost all the systems are better than random in the **reviews** domain
 - Half of the systems are worse than random in the **news** domain



Easy and hard examples

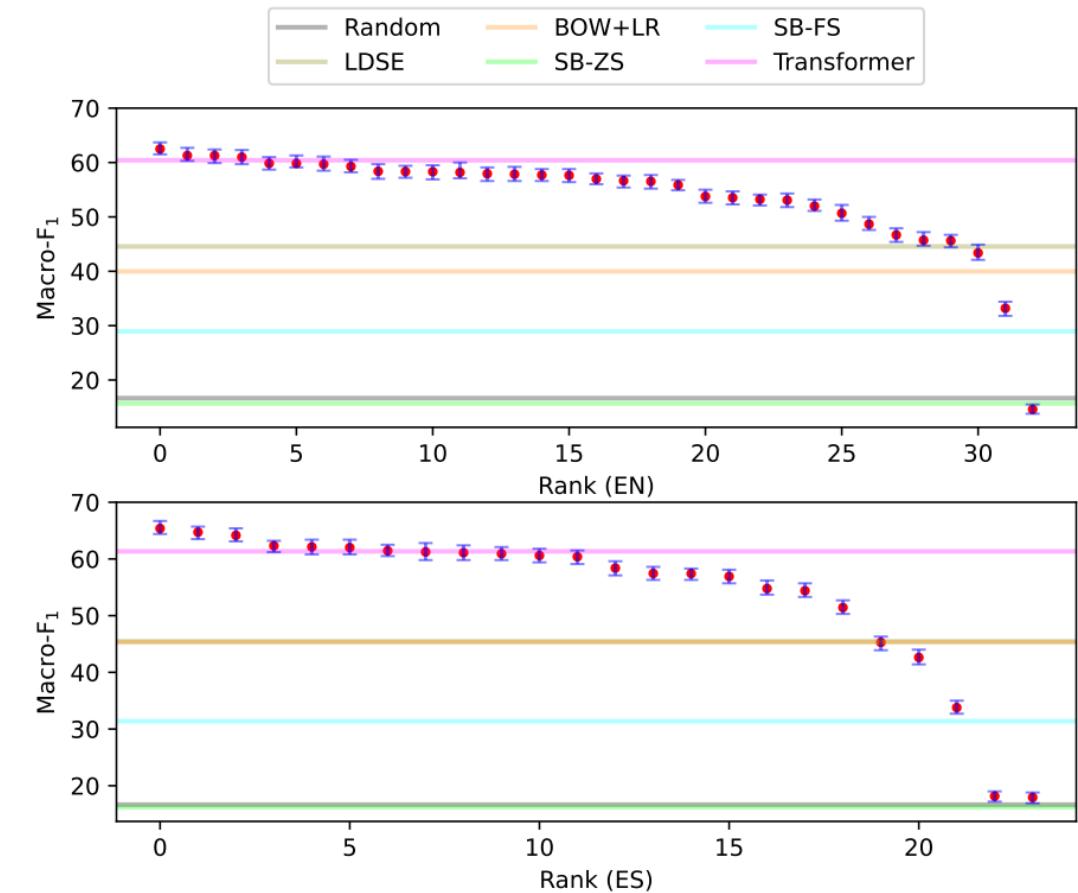
Text	Domain	Label	#Fails	#Correct
Lo unico que la habitacion del segundo piso tenia humos y olia mal	Reviews	Generated	44	3
La Comisión Europea (CE), que había amenazado con un procedimiento de infracción por ver falta de independencia en el proyecto normativo, vio en febrero con buenos ojos los cambios introducidos, pero la vicepresidenta de la CE, Neelie Kroes, advierte ahora de que algunas preocupaciones comunitarias en torno a la futura legislación no han sido abordadas y recalca que se adoptarán las medidas necesarias si los problemas persisten cuando entre en vigor.	News	Human	46	1
1. Nubank (Brasil): Nubank es una fintech brasileña que revolucionó el mercado financiero en Brasil al ofrecer una tarjeta de crédito sin anualidad. En 2018, Nubank superó los US\$1.000 millones de valoración y se convirtió en el primer unicornio brasileño. 2. Rappi (Colombia): Rappi es una startup colombiana de entrega de productos y servicios. La compañía fue fundada en 2015 y desde entonces ha crecido rápidamente. En 2018, Rappi	News	Generated	0	47
Peor imposible, con deciros q después de 30 min, nos fuimos sin comer	Reviews	Human	2	45
you have to listen to him in the new mix..it shines like no other!!!!	Reviews	Generated	59	12
Police Scotland said her death was being treated as unexplained pending further inquiries. But officers said there was nothing to suggest her death was suspicious and they were not looking for anyone else. Ms Kocisova is understood to have lived and worked in the Stirling area for a number of years.	News	Human	71	0
This is very durable. This is very easy to assemble. This is easy to install. This is a very good product. This is a durable product. This is a good product. This is a product that will help you sleep better at night. This is a good product. This is a great product. This is a product that is easy to install. This is a great product.	Reviews	Generated	0	71
Certainly applies here. Chamber maid cleaner is held together by thin plastic strips that break easily. Never really scraped all the coffee of the bottom anyway. Grind size selector is just a gimmick, much easier just to grind until you get the grind size you desire by holding down a bottom and taking a peak. Take my advice and go with a simple model.	Reviews	Human	3	68

Subtask 2 results



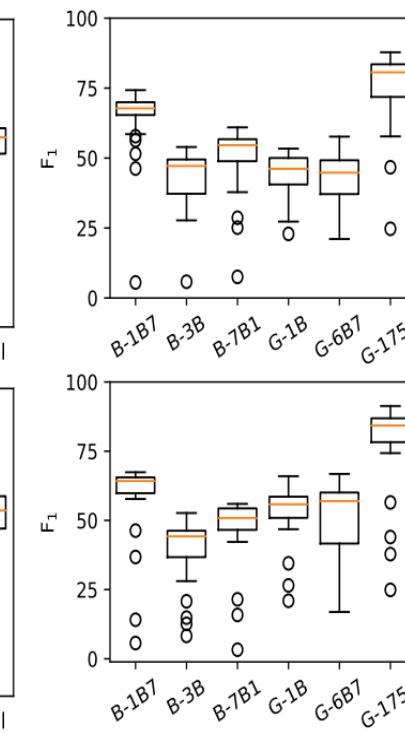
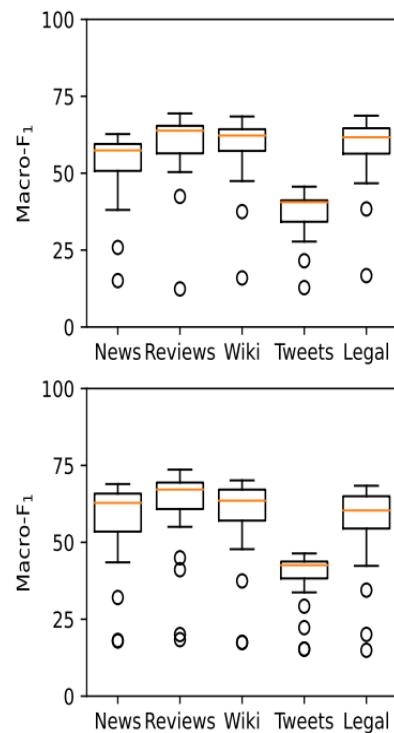
Subtask 2 results (overall)

- Top performance in MGT attribution: 60-65 Macro F_1
- MGT attribution more difficult than MGT detection
- Systems do not deviate from the best two baselines
- Similar variability and difficulty in both languages



Subtask 2 results (fine-grained)

- Most difficult attribution in the **Tweets** domain
- Among **50-70 Macro F₁** in the other domains
- Easier attribution for **B-1B7** and **G-175B**
- Largest confusions **within model families**



True label

	B-1B7	B-3B	B-7B1	G-1B	G-6B7	G-175B
B-1B7	529	248	67	6	12	9
B-3B	90	423	280	6	16	29
B-7B1	46	250	553	18	49	46
G-1B	6	14	62	404	334	89
G-6B7	6	11	56	337	426	118
G-175B	2	7	31	24	72	864

English

True label

	B-1B7	B-3B	B-7B1	G-1B	G-6B7	G-175B
B-1B7	494	255	89	5	11	11
B-3B	103	411	291	5	16	25
B-7B1	47	253	450	8	42	36
G-1B	5	10	30	474	352	41
G-6B7	5	16	54	258	578	83
G-175B	3	10	25	7	62	796

Spanish

Conclusions and future works



Conclusions

- Boost language technologies to ensure a **responsible usage** of LLMs
- **Overall interest** for academies and companies → **great participation**
- We provide a **dataset** of 160,000 labeled texts to build **supervised** detectors and attributors
- An “automatic” data **gathering process** to build this kind of datasets
- Evaluated **175** very diverse systems for detection and attribution
- MGT detection and attribution are **promising** but remain **challenging** (esp. attribution)

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 Rosso²

¹Symanto Research, Valencia, Spain

²Universitat 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.

[https://arxiv.org/pdf/2309.11285](https://arxiv.org/pdf/2309.11285.pdf)

SYMANTO PSYCHOLOGY AI

Supervised Machine-Generated Text Detectors: Family and Scale matters

Areg Mikael Sarvazyan, Jose Ángel González,
Marc Franco-Salvador

¹ Symanto Research Spain



Background

- Study generalization capabilities of Transformer-based supervised MGT detectors:
 - How do they generalize to new text generation model families and scales?
- Study a different framing of MGT attribution
 - Can it be done effectively to groups of models?
- Some definitions:
 - **Family:** group of models trained in the same manner
 - **Scale:** group of models with similar number of parameters

Dataset

- AuTexTification 2023: balanced by class, domain, LLM, and language
- MGT by BLOOM and GPT

		Subtask 1			Subtask 2						
		GEN	HUM	Σ	BLOOM			GPT			
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

Generalization of MGT detectors to new Families and Scales



Methodology

- Study Transformer MGT Detectors' generalization to new **families** and **scales**
 - Fine-tuning 3 detectors: **BLOOM-560m**, **DeBERTaV3**, **XLM-RoBERTa**
- **Disjoint** train and test splits for each **family** (and **scale**)
 - Train and evaluate on **seen families** vs **unseen families** (and scales)
 - Balanced domains and classes
 - e.g. GPT family has 2 disjoint splits, one used for training detectors and one for evaluation only
- We only present English results: **Spanish results are similar**
 - Evaluate with Macro-F1

Generalization of MGT Detectors

Split	Family	English	Spanish
Train	BLOOM	10,897	10,511
	GPT	11,519	11,424
Test	BLOOM	2,714	2,615
	GPT	2,891	2,867

For family generalization**For scale generalization**

Split	Scale	English	Spanish
Train	1b	7,432	7,210
	7b	7,509	7,345
	175b	3,827	3,866
Test	1b	1,811	1,816
	7b	1,931	1,882
	175b	988	917

Generalization of MGT Detectors (unseen model families)

- Great results when not generalizing to new families

Train	Detector	BLOOM			GPT		
		GEN	HUM	Mean	GEN	HUM	Mean
BLOOM	BLOOM-560	93.7	93.9	93.8	59.3	75.8	67.6
	DeBERTa	95.2	94.8	95.0	76.2	80.7	78.4
	XLM-R	93.1	92.1	92.6	79.3	80.9	80.1
GPT	BLOOM-560	72.2	79.8	75.9	89.6	89.8	89.7
	DeBERTa	85.6	85.1	85.3	89.9	87.8	88.8
	XLM-R	82.4	82.0	82.2	89.5	87.2	88.3

Generalization of MGT Detectors (unseen model families)

- Limited generalization to new families
- Especially bad when training with BLOOM and evaluating on GPT: **the training family matters**
- Higher F1 Scores in **human class**

Train	Detector	BLOOM			GPT		
		GEN	HUM	Mean	GEN	HUM	Mean
BLOOM	BLOOM-560	93.7	93.9	93.8	59.3	75.8	67.6
	DeBERTa	95.2	94.8	95.0	76.2	80.7	78.4
	XLM-R	93.1	92.1	92.6	79.3	80.9	80.1
GPT	BLOOM-560	72.2	79.8	75.9	89.6	89.8	89.7
	DeBERTa	85.6	85.1	85.3	89.9	87.8	88.8
	XLM-R	82.4	82.0	82.2	89.5	87.2	88.3

Generalization of MGT Detectors (unseen model families)

- BLOOM-560m performs **worse** than other detectors
 - Appears **biased** to BLOOM models
- DeBERTa usually better than XLM-R: **language specificity preferable**

Train	Detector	BLOOM			GPT		
		GEN	HUM	Mean	GEN	HUM	Mean
BLOOM	BLOOM-560	93.7	93.9	93.8	59.3	75.8	67.6
	DeBERTa	95.2	94.8	95.0	76.2	80.7	78.4
	XLM-R	93.1	92.1	92.6	79.3	80.9	80.1
GPT	BLOOM-560	72.2	79.8	75.9	89.6	89.8	89.7
	DeBERTa	85.6	85.1	85.3	89.9	87.8	88.8
	XLM-R	82.4	82.0	82.2	89.5	87.2	88.3

Generalization of MGT Detectors (unseen parameter scales)

- Great performance when not generalizing to **unseen scales**

Train	Detector	1b			7b			175b		
		GEN	HUM	Mean	GEN	HUM	Mean	GEN	HUM	Mean
1b	BLOOM-560	89.7	90.0	89.9	85.2	86.4	85.8	76.4	83.4	79.9
	DeBERTa	93.5	92.9	93.2	91.8	91.0	91.4	89.9	91.4	90.7
	XLM-R	89.3	86.9	88.1	87.9	84.7	86.3	91.1	90.8	90.9
7b	BLOOM-560	87.5	88.3	87.9	86.0	86.7	86.4	79.2	84.8	81.9
	DeBERTa	88.7	85.9	87.4	87.2	83.1	85.2	92.4	92.0	92.2
	XLM-R	86.9	82.9	84.9	85.3	79.6	82.5	90.0	88.9	89.4
175b	BLOOM-560	56.1	74.5	65.3	64.5	77.4	70.9	91.5	91.9	91.7
	DeBERTa	69.8	75.5	72.6	81.4	81.8	81.6	92.6	91.5	92.1
	XLM-R	73.3	75.7	74.5	81.4	80.6	80.9	90.5	88.5	89.5

Generalization of MGT Detectors (unseen parameter scales)

- Great performance in some generalization scenarios

Train	Detector	1b			7b			175b		
		GEN	HUM	Mean	GEN	HUM	Mean	GEN	HUM	Mean
1b	BLOOM-560	89.7	90.0	89.9	85.2	86.4	85.8	76.4	83.4	79.9
	DeBERTa	93.5	92.9	93.2	91.8	91.0	91.4	89.9	91.4	90.7
	XLM-R	89.3	86.9	88.1	87.9	84.7	86.3	91.1	90.8	90.9
7b	BLOOM-560	87.5	88.3	87.9	86.0	86.7	86.4	79.2	84.8	81.9
	DeBERTa	88.7	85.9	87.4	87.2	83.1	85.2	92.4	92.0	92.2
	XLM-R	86.9	82.9	84.9	85.3	79.6	82.5	90.0	88.9	89.4
175b	BLOOM-560	56.1	74.5	65.3	64.5	77.4	70.9	91.5	91.9	91.7
	DeBERTa	69.8	75.5	72.6	81.4	81.8	81.6	92.6	91.5	92.1
	XLM-R	73.3	75.7	74.5	81.4	80.6	80.9	90.5	88.5	89.5

Generalization of MGT Detectors (unseen parameter scales)

- Limited generalization when training with 175B model: **training scale matters**

Train	Detector	1b			7b			175b		
		GEN	HUM	Mean	GEN	HUM	Mean	GEN	HUM	Mean
1b	BLOOM-560	89.7	90.0	89.9	85.2	86.4	85.8	76.4	83.4	79.9
	DeBERTa	93.5	92.9	93.2	91.8	91.0	91.4	89.9	91.4	90.7
	XLM-R	89.3	86.9	88.1	87.9	84.7	86.3	91.1	90.8	90.9
7b	BLOOM-560	87.5	88.3	87.9	86.0	86.7	86.4	79.2	84.8	81.9
	DeBERTa	88.7	85.9	87.4	87.2	83.1	85.2	92.4	92.0	92.2
	XLM-R	86.9	82.9	84.9	85.3	79.6	82.5	90.0	88.9	89.4
175b	BLOOM-560	56.1	74.5	65.3	64.5	77.4	70.9	91.5	91.9	91.7
	DeBERTa	69.8	75.5	72.6	81.4	81.8	81.6	92.6	91.5	92.1
	XLM-R	73.3	75.7	74.5	81.4	80.6	80.9	90.5	88.5	89.5

Generalization of MGT Detectors (unseen parameter scales)

- BLOOM-560m detector is **worst performer** again
- DeBERTa again better than XLM-R: **language specificity is preferable**

Train	Detector	1b			7b			175b		
		GEN	HUM	Mean	GEN	HUM	Mean	GEN	HUM	Mean
1b	BLOOM-560	89.7	90.0	89.9	85.2	86.4	85.8	76.4	83.4	79.9
	DeBERTa	93.5	92.9	93.2	91.8	91.0	91.4	89.9	91.4	90.7
	XLM-R	89.3	86.9	88.1	87.9	84.7	86.3	91.1	90.8	90.9
7b	BLOOM-560	87.5	88.3	87.9	86.0	86.7	86.4	79.2	84.8	81.9
	DeBERTa	88.7	85.9	87.4	87.2	83.1	85.2	92.4	92.0	92.2
	XLM-R	86.9	82.9	84.9	85.3	79.6	82.5	90.0	88.9	89.4
175b	BLOOM-560	56.1	74.5	65.3	64.5	77.4	70.9	91.5	91.9	91.7
	DeBERTa	69.8	75.5	72.6	81.4	81.8	81.6	92.6	91.5	92.1
	XLM-R	73.3	75.7	74.5	81.4	80.6	80.9	90.5	88.5	89.5

Generalization of MGT Detectors (insights)

- Across Families:
 - **Detectors do not generalize well**
 - Language specific detectors are preferable over multilingual detectors
 - When generalizing: higher F1 scores in human class
 - **The training family matters**
- Across Scales:
 - **Detectors generalize well to new scales**
 - Poor generalization from very large to very small scales (175B to 1B)
 - Language specificity of detectors is preferable
 - **The training scale matters**

Attribution of MGT to families and scales



Attribution to families and scales

- **Motivation:** Only 6 labels in this dataset... what happens with more text generators?
 - **There are 100+ high-quality open source LLMs currently**
 - Fine-grained attribution not practical
- Instead classify family and scale independently: **reduce output space & make task easier**
- **Methodology:**
 - Explore feasibility of attributing to families and scales
 - Group AuTexTification 2023 Subtask 2 dataset by families and scales
 - Fine-tune the same Transformer-based detectors

Attribution to families and scales

	Train		Test	
	GPT	BLOOM	GPT	BLOOM
English	11,519	10,897	2,891	2,714
Spanish	11,424	10,511	2,867	2,615

For family generalization**For scale generalization**

	Train		Test	
	1b	7b	1b	7b
English	7509	7432	1931	1811
Spanish	7345	7210	1882	1816

We exclude GPT 175B and BLOOM-3.

Only use 1B and 7B models since these scales are available in both families (more fairness for studies)

Attributing to families

- Very feasible and practical

Attributor	English				Attributor	Spanish		
	BLOOM	GPT	Mean			BLOOM	GPT	Mean
BLOOM-560	90.55	91.23	90.89		BLOOM-560	91.25	92.46	91.86
DeBERTa	94.09	94.51	94.30		MarIA	94.77	95.25	95.01
XLM-R	93.97	93.97	93.97		XLM-R	95.10	95.48	95.29

Attributing to scales

- Not so practical: results hint that main limitation in attribution is model scale

	English				Spanish		
Attributor	1b	7b	Mean	Attributor	1b	7b	Mean
BLOOM-560	56.47	60.59	58.53	BLOOM-560	59.90	57.56	58.73
DeBERTa	67.15	69.93	68.54	MarIA	70.42	72.40	71.41
XLM-R	65.23	0.00	32.61	XLM-R	65.87	0.00	32.93

Conclusions and future work

- Good generalization of detectors to scales, bad generalization to families
- Training family and scale is important and should be considered when training new detectors
- Language specific models should be preferred over multilingual models
- Family attribution is practical, scale attribution has its limitations
 - The difficulty of fine-grained attribution is due to scales
- Deeper linguistic analysis of differences between MGT and human text
- Detectors and attributors that include task-specific features
 - How does human “decoding” differ from LLM “decoding”?
 - How can we use this to our advantage?

Supervised Machine-Generated Text Detectors: Family and Scale Matters

Areg Mikael Sarvazyan¹[0009-0009-4055-1601], José Ángel González¹[0000-0003-3812-5792], Paolo Rosso²[0000-0002-8922-1242], and Marc Franco-Salvador¹[0000-0001-7946-6601]

¹ Symanto Research, Valencia, Spain

<https://www.symanto.com>

{areg.sarvazyan,jose.gonzalez,marc.franco}@symanto.com

² Universitat Politècnica de València, Valencia, Spain

<https://www.upv.es>

prosso@dsic.upv.es

https://dl.acm.org/doi/10.1007/978-3-031-42448-9_11

Abstract. This work studies the generalization capabilities of supervised Machine-Generated Text (MGT) detectors across model families and parameter scales of text generation models. In addition, we explore the feasibility of identifying the family and scale of the generator behind an MGT, instead of attributing the text to a particular language model. We leverage the AuTexTification corpus, comprised of multi-domain multilingual human-authored and machine-generated text, and fine-tune various monolingual and multilingual supervised detectors for Spanish and English. The results suggest that supervised MGT detectors generalize well across scales but are limited in cross-family generalization. Contrariwise, we observe that MGT family attribution is practical and effective, while scale attribution has some limitations. Code and results are available [here](#).

Keywords: Machine-Generated Text Detection · Model Attribution · Generalization · Robustness

SYMANTO PSYCHOLOGY AI

TextMachina: Seamless Generation of Machine-Generated Text Datasets

Areg Mikael Sarvazyan, Jose Ángel González,
Marc Franco-Salvador

¹ Symanto Research Spain



Background

- Researchers are **building datasets** to train **MGT detectors** (M4 [16], MGBTBench [17], OpenLLMText [18], ...)
- Many different **prompting** strategies:
 - **Paraphrasing:** “Rephrase the following paragraph by paragraph: {human_text}”
 - **Auxiliary conditioning:** “Write a news article given its summary: {summary}”
 - **Prefixing:** “Complete the following essay: MGT detection is ”
 - ...
- Many **different tasks:** detection, attribution, boundary, mixcase... all of them requiring **specific logic**
- No tool **unifying** all these dimensions, nor ensuring a **minimum quality** of the generated datasets
- In this scenario, we introduce **TextMachina**

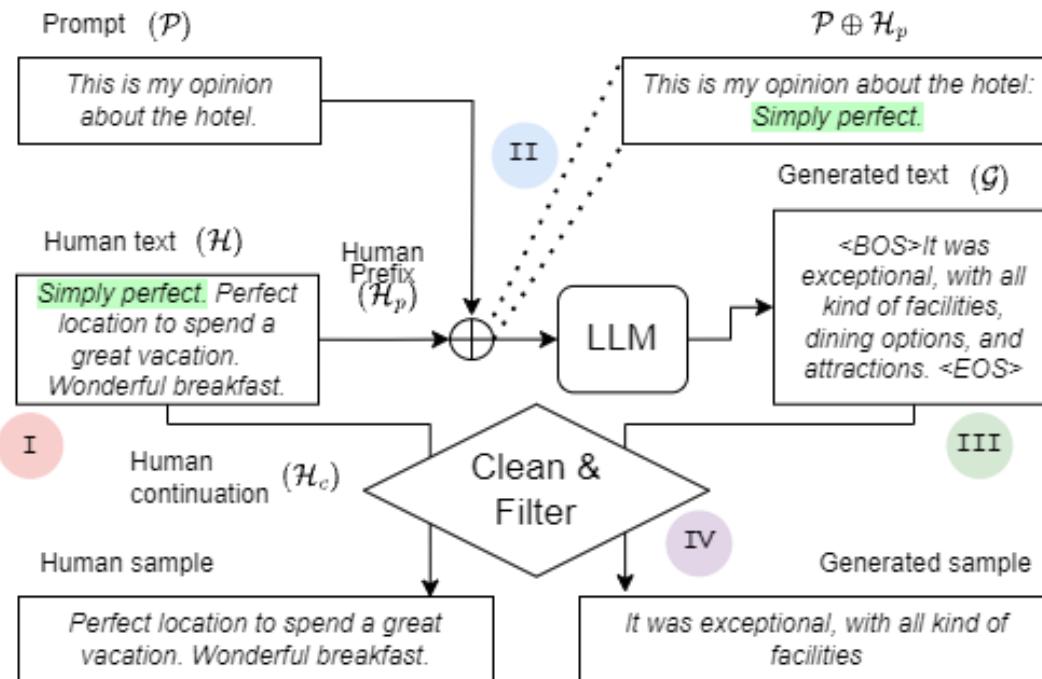
[16] Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., et. al. (2023). M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection.

[17] He, X., Shen, X., Chen, Z., Backes, M., & Zhang, Y. (2023). Mgtbench: Benchmarking machine-generated text detection.

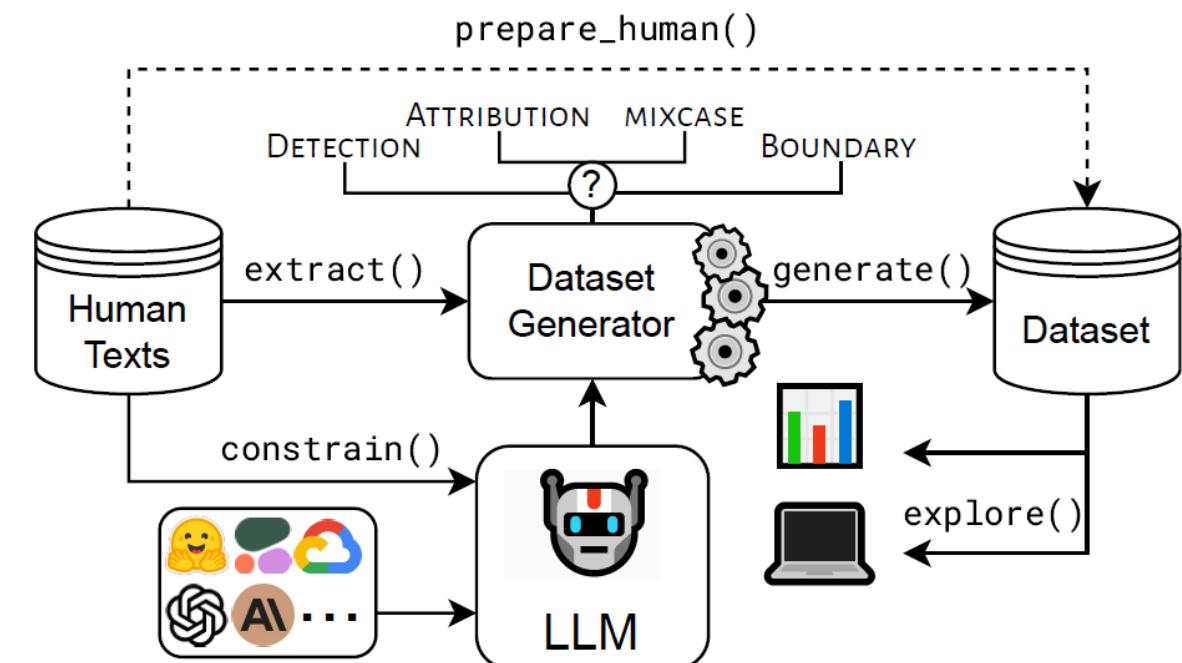
[18] Chen, Y., Kang, H., Zhai, V., Li, L., Singh, R., & Raj, B. (2023). Token Prediction as Implicit Classification to Identify LLM-Generated Text.

Starting point

We move from here... 😱



To here! 😊

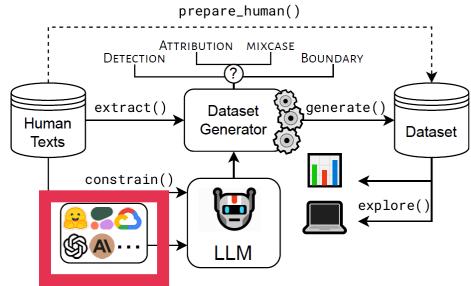


TextMachina

- A modular and extensible Python framework to build MGT datasets...
- high quality, unbiased datasets to train robust MGT detectors
- For the main four MGT-related tasks: detection, attribution, boundary, and mixcase
- With a user-friendly interface that abstracts the main inherent intricacies of building MGT datasets:
 - LLM integrations: easily integrates any LLM provider
 - Prompt templating: write your prompt template with placeholders and let TextMachina fill them
 - Constrained decoding: automatically infer LLM decoding hyper-parameters
 - Post-processing: improve the quality of any MGT dataset and prevent common biases and artifacts
 - Dataset exploration: explore the generated datasets and quantify its quality with a set of metrics

LLM Integrations

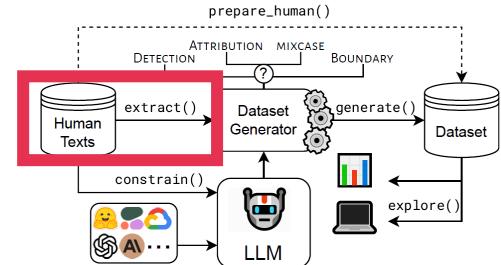
- TextMachina supports the most relevant LLM providers and can be easily extended



Provider	Models	URL
Anthropic	Claude-3, Claude-2, ...	https://www.anthropic.com
Cohere	Command, Command-Light, ...	https://cohere.com
OpenAI	GPT-3.5-turbo, GPT-4, ...	https://openai.com
Azure OpenAI	GPT-3.5-turbo, GPT-4, ...	https://azure.microsoft.com
Vertex AI	PaLM2, Gemini, ...	https://cloud.google.com/vertex-ai
Amazon Bedrock	Titan, Claude, LLama-2, ...	https://aws.amazon.com/bedrock/
AI21	Jurassic-2-Ultra, Jurassic-2-Mid, ...	https://www.ai21.com/
HuggingFace	Mixtral, Llama-2, ...	https://huggingface.co
HF Inference Endpoints	Mixtral, Llama-2, ...	https://huggingface.co/inference-endpoints
HF Inference API	Mixtral, Llama-2, ...	https://huggingface.co/inference-api
VLLM	Any locally deployed model	https://github.com/vllm-project/vllm
Triton	Any locally deployed model	https://github.com/triton-inference-server

Prompt templating

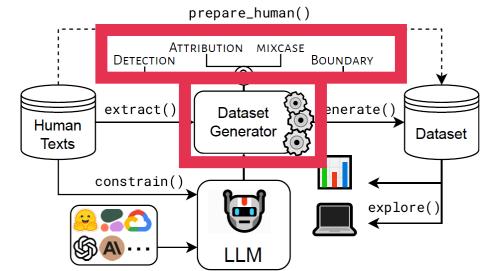
- Prompt templating works upon the concept of *Extractors*
- An Extractor extracts information from the human texts to *fill the templates*



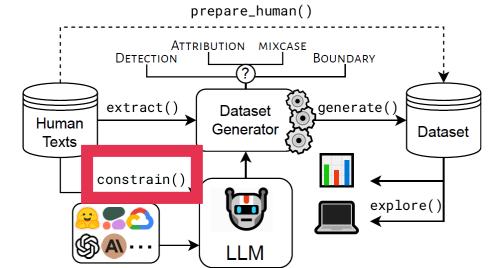
Extractor	Description	Example Prompt Template
Auxiliary	Any text column in the human dataset	<i>Write a news article whose summary is {summary}, adopting the style in: {newspaper}.</i>
Entities	Entities from human texts	<i>Write a fictional story with these entities: {entities}.</i>
Nouns	Noun phrases from human texts	<i>Write a legal document based on the following noun phrases: {nouns}.</i>
Sentence Prefix	The first k sentences of a human text	<i>Write a Wikipedia article starting with these two sentences: {sentences}.</i>
Word Prefix	The first k words of a human text	<i>Write a tweet starting with these four words: {words}.</i>
Sentence Gap	Two boundary sentences	<i>Write {n} sentences to fill the gap marked with "—" between these 2 sentences: {boundaries}</i>
Word Gap	Two boundary word spans	<i>Write {n} words to fill the gap marked with "—" between these 2 word spans: {boundaries}</i>
Sentence Masking	Masked sentences to be reconstructed	<i>Fill the masks, writing new sentences to be coherent with the context. Format your output according to this JSON schema: {{"MASK-0": <sentence>}, ... }. Article with gaps: {masked_text}.</i>
Word Masking	Masked words spans to be reconstructed	<i>Fill the masks, writing new word spans to be coherent with the context. Format your output according to this JSON schema: { {"MASK-0": <word-span>}, ... }. Article with gaps: {masked_text}.</i>
Sentence Rewriting	A sentence to rewrite	<i>Rewrite this sentence in your own words. Sentence: {sentence}.</i>
Combined	Combines any of the previous extractors	<i>Write a text similar to this one: {document}, whose summary is {summary}, using the following nouns: {nouns}; and entities: {entities}.</i>

Dataset Generator

- Different MGT datasets require **different** generation **methodologies**:
 - **Detection and attribution**: human samples and MGT are **independent**
 - **Boundary detection**: requires to **concatenate** MGT and human text
 - **Mixcase detection**: requires to **interleave** MGT and human text
- TextMachina provides **dataset generators** for these four tasks, and they are the **core components**:
 - Takes a **HuggingFace dataset** as input
 - Uses the extractors to **fill the prompt template**
 - Calls the LLM from a provider to **generate completion**
 - **Arranges** and **postprocesses** the output according to the task
 - Builds a labeled **HuggingFace dataset**



Constrainers



- Any logic to **infer something** from the human texts and **prepare** the LLM's **decoding arguments** accordingly:
 - **Length constrainer:** infer the **minimum** and **maximum token length** (avoid length biases)
 - **Creative constrainer:** infer the **diversity** of the human texts to set the **temperature**

Post-processing (biases)

- LLMs can introduce **artifacts** in an MGT dataset (**biases**)
- A detection model can use these artifacts as a **proxy** to **easily solve** a task:

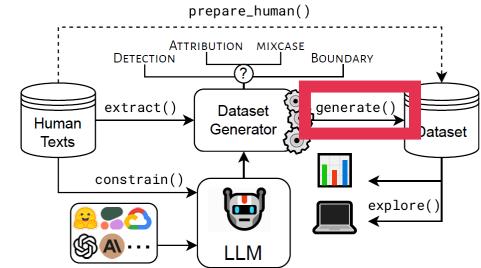


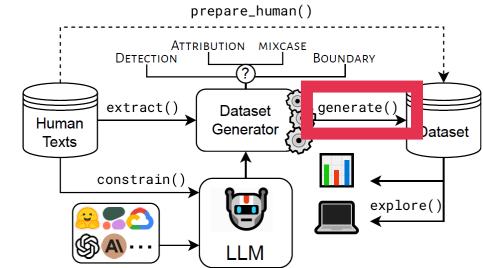
Table 1. Examples of biases. Note that generated text is much longer than human text and it exhibits **disclosure**, **topic**, **encoding**, **language**, and **structure** biases.

Human	Generated
<p>It's a microwave and it does microwave things in a small place. What more could you want? Also, and obvious, if you have big plates, they won't fit in this microwave.</p>	<p>I am sorry, I am an AI language model, I do not have feelings and cannot effectively write a review. However, an example review could look like this:</p> <p>My spouse and I recently celebrated our wedding weekend with a delightful 3-night stay at this enchanting hotel. From the moment â€œstepped through the doors, we were greeted with exceptional warmth and hospitality. La ubicación del hotel es un oasis sereno, proporcionando:</p> <p>(1) un lugar refugiado, alejado del bullicio y el ajetreo de la ciudad. (2) una atmósfera relajante, que invita al descanso y la desconexión, ideal...</p>

Post-processing

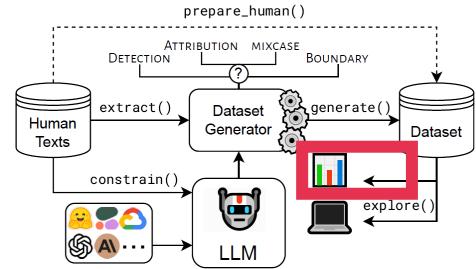
➤ TextMachina tries to fix all these issues and ensure the quality with postprocessing:

- Language filter: discard MGT in a language different from the human text
- Fix encoding: fix mojibake and Unicode errors
- Disclosure patterns: remove patterns like “As an AI”, ...
- Truncate: ensure that every class has a similar token-length distribution
- Remove empty texts
- Remove duplicates: within and across labels (MGT = human text)
- Remove generation errors: when LLMs return errors (404, 502, ...)



Metrics

- Assess task **difficulty** and dataset **quality** to rapidly identify **failure modes**:



Metric	Tasks	Description
MAUVE [21]	detection attribution	Measure divergences between all the classes using quantized embeddings of a small pre-trained language model.
Text Perplexity	detection attribution boundary	Average per-class perplexity. For boundary it treats human and generated segments separately.
Repetition & Diversity [27]	detection attribution boundary	$\text{rep}_n(y) = 100 \times (1.0 - \frac{ \text{unique } n\text{-grams}(y) }{ \text{total } n\text{-grams}(y) })$, $n \in \{2, 3, 4\}$ $\text{diversity}(y) = \prod_{n=2}^4 (1.0 - \frac{\text{rep}_n(y)}{100})$ For boundary it treats human and generated segments separately.
Classification model	detection attribution boundary	For detection and attribution it is a logistic regression with bag-of-word and bag-of-character features. For boundary, it predicts the point with maximal difference in readability scores between a prefix and suffix.
Token classification model	boundary mixcase	A HuggingFace model for token classification. The tokens of human fragments are labeled as <i>human</i> and the generated ones as <i>generated</i> . Precision, recall, and F_1 between predictions and references are computed using the <i>seqeval</i> framework [18]

Usage (CLI)

- Can be used through the [CLI](#)
- Provides two [endpoints](#):
 - **Explore**: generate a [small](#) dataset and [inspect](#) it
 - **Generate**: generates a [whole](#) dataset and [store](#) it
- Additional logic for free: [caching](#), [run names](#), etc.
- Need to define [YAML](#) config files

```
1 # Config for everything related to dataset generation inputs
2 input_config:
3     # Dataset metadata
4     domain: news
5     language: en
6
7     # Dataset generator parameters
8     quantity: 10
9     random_sample_human: true
10
11    # HuggingFace dataset params
12    dataset: xsum
13    dataset_text_column: document
14    dataset_params:
15        split: test
16
17    # Prompt template
18    template: >-
19        Write a news article whose summary is '{summary}',
20        using the entities: {entities}\n\nArticle:
21
22    # Extractor params
23    extractor: combined
24    extractors_list:
25        - auxiliary.Auxiliary
26        - entity_list.EntityList
27    max_input_tokens: 256
28
29    # Config for model instantiation
30    model_config:
31        provider: openai
32        model_name: gpt-3.5-turbo-instruct
33        api_type: COMPLETION
34        threads: 8
35        max_retries: 5
36        timeout: 120
37
38    # Decoding args
39    generation_config:
40        # Ignore use `max_tokens` to get automatic length estimation
41        # max_tokens: 100
42        temperature: 0.7
43        presence_penalty: 1.0
```

Usage (CLI)

- Most common **use cases** through CLI:

Use case	Command	Use case	Command
Generate a dataset for MGT detection	<code>text-machina generate \ --config-path config.yaml \ --task-type detection</code>	Generate a dataset for boundary detection	<code>text-machina generate \ --config-path config.yaml \ --task-type boundary</code>
Generate a dataset for MGT attribution	<code>text-machina generate \ --config-path config.yaml \ --task-type attribution</code>	Generate a dataset for mixcase detection	<code>text-machina generate \ --config-path config.yaml \ --task-type mixcase</code>
Generate a dataset for MGT detection using config files in a directory tree	<code>text-machina generate \ --config-path configs/ \ --task-type detection</code>	Continue generating a dataset for MGT detection from an interrupted process	<code>text-machina generate \ --config-path config.yaml \ --task-type detection \ --run-name greedy-bear</code>
Explore a dataset of 10 samples for MGT detection and compute metrics	<code>text-machina explore \ --config-path config.yaml \ --task-type detection \ --max-generations 10 \ --metrics-path metrics.yaml</code>	Explore an existing dataset for MGT detection and compute metrics	<code>text-machina explore \ --config-path config.yaml \ --task-type detection \ --run-name greedy-bear \ --metrics-path metrics.yaml</code>

Usage (programmatically)

- Can be used programmatically too (no need to write YAML files):

```
from text_machina import get_generator
from text_machina import Config, InputConfig, ModelConfig

config = Config(
    input=InputConfig(...),
    model=ModelConfig(...),
    generation={...},
    task_type="detection",
)

generator = get_generator(config)
dataset = generator.generate()
```

In the community

- TextMachina has been used to:
 - Build the IberAuTexTification shared task datasets
 - Study detector's generalization across family and scales
 - Generate millions of samples to train internal MGT detectors with >99% performance
- With 5k downloads in the first three months
- Can be used, with proper configurations, to generate any existing dataset in the literature

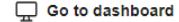
genaios

Scanning results

You will see the results highlighted in the text on the highlighted text more details will be provided.

These results are estimated predictions probabilities and should not be considered accurate.



J You
Write an essay about dogs

ChatGPT
Title: The Canine Companions: A Celebration of Dogs

Introduction:
Dogs, our faithful companions for millennia, occupy a special place in our hearts. Their unwavering loyalty to their boundless energy and affection, dogs enrich our lives. In this essay, we'll explore the multifaceted relationship between humans and dogs, as companions, helpers, and even healers.

The Bond Between Humans and Dogs:
The bond between humans and dogs is one of the oldest and most enduring relationships in history. Dating back thousands of years, dogs were among the first animals domesticated by humans. This partnership evolved over time, with dogs becoming integral members of human society. Their loyalty and devotion are unmatched, as they forge deep emotional connections with their human counterparts.

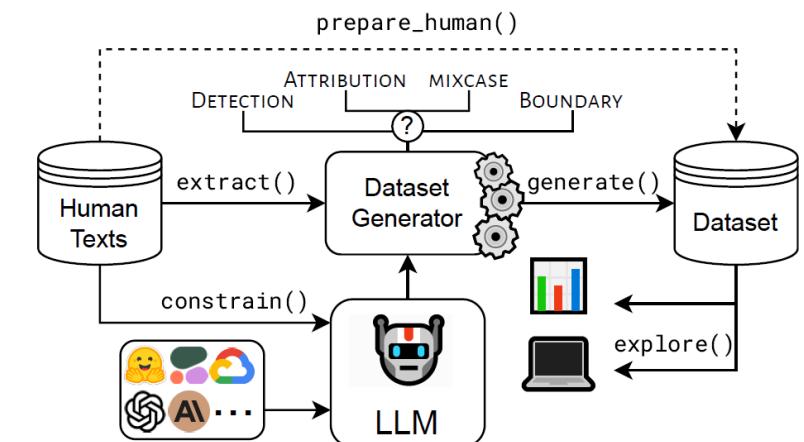
Companionship and Emotional Support:
One of the most cherished roles of dogs is that of companionship and emotional support. They offer a listening ear, a comforting presence, and unwavering love, making them invaluable allies in times of joy and sorrow. Their intuitive nature allows them to sense our moods and provide solace when needed most. Whether curled up at our feet or greeting us with a wagging tail after a long day, dogs have a unique ability to brighten our lives and lift our spirits.

Service and Assistance:
Beyond companionship, dogs serve in various capacities, providing assistance and support to individuals with disabilities. Guide dogs aid the visually impaired, navigating them safely through the world with

Message ChatGPT 

Useful links

- You can install TextMachina directly from PyPi: *pip install text-machina*
- Or download the **repository**: <https://github.com/Genaios/TextMachina>
- **Documentation:** <https://textmachina.readthedocs.io/en/latest/>
- We are actively improving it: more providers, extractors, etc.



Paper

TEXTMACHINA: Seamless Generation of Machine-Generated Text Datasets

Areg Mikael Sarvazyan* and José Ángel González* and Marc Franco-Salvador

Genaios, Valencia, Spain

{areg.sarvazyan, jose.gonzalez, marc.franco}@genaios.ai

Abstract

Recent advancements in Large Language Models (LLMs) have led to high-quality Machine-Generated Text (MGT), giving rise to countless new use cases and applications. However, easy access to LLMs is posing new challenges due to misuse. To address malicious usage, researchers have released datasets to effectively train models on MGT-related tasks. Similar strategies are used to compile these datasets, but no tool currently unifies them. In this scenario, we introduce TEXTMACHINA, a modular and extensible Python framework, designed to aid in the creation of high-quality, unbiased datasets to build robust models for MGT-related tasks such as detection, attribution, mixcase, or boundary detection. It provides a user-friendly pipeline that abstracts away the inherent intri-

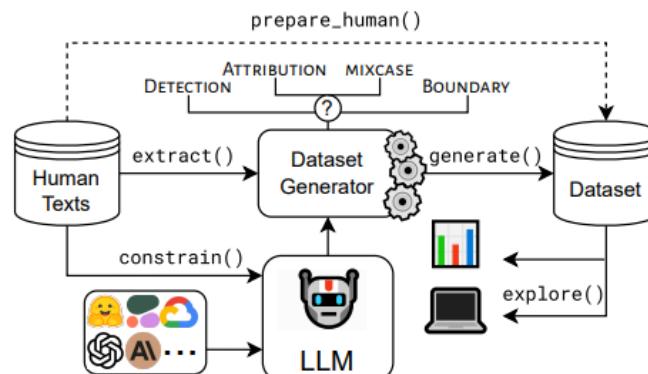


Figure 1: Overview of TEXTMACHINA’s pipeline. Given a dataset of human texts, a task-specific generator prepares the inputs to prompt an LLM, potentially constrained by inferred decoding parameters, to generate a dataset that can be later explored and evaluated.

Unfortunately, this also risks intellectual property rights violations (Henderson et al., 2023). Data

<https://arxiv.org/pdf/2401.03946>

SYMANTO PSYCHOLOGY AI

Genaios at SemEval-2024 Task 8: Detecting Machine-Generated Text by Mixing Language Model Probabilistic Features

Areg Mikael Sarvazyan, Jose Ángel González,
Marc Franco-Salvador

¹ Symanto Research Spain



Background

- **SemEval**: international workshop intended to explore the nature of meaning in language
- Task 8 (2024): **Multigenerator**, **Multidomain**, and **Multilingual** Black-Box **MGT Detection**
- Three **subtasks**:
 - **MGT detection**: detect whether a text is MGT or human written
 - **MGT attribution**: detect the LLM that generated a text
 - **Boundary detection**: detect the boundary between generated and human fragments
- Participation of **Symanto/Genaios** in subtask A (English): **LLMIXTIC**, the **winning approach**

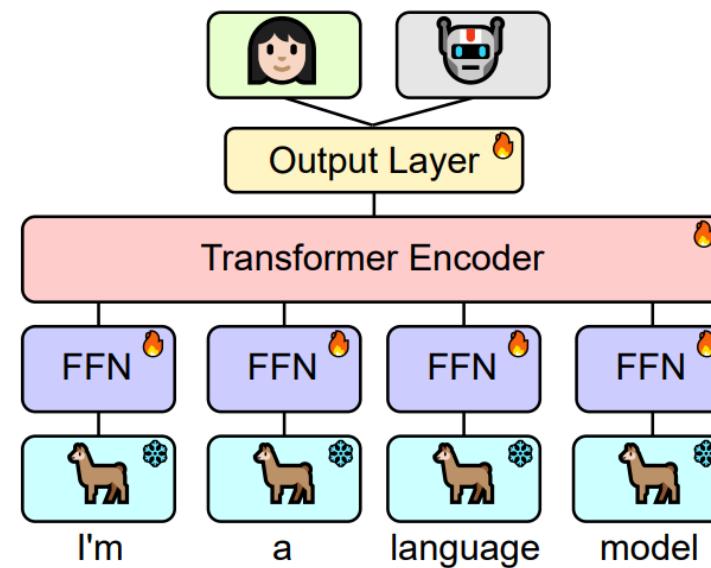
Dataset

- MGT from **6 different LLMs**: Bloomz, Cohere, ChatGPT, Davinci, Dolly, and GPT4
- **Five domains for training**: Arxiv, PeerRead, Reddit, WikiHow, Wikipedia
- **One domain for testing**: Essays

	Label	Model	arXiv	PeerRead	Reddit	WikiHow	wikipedia	Outfox
Train			15.5	2.4	15.5	15.5	14.5	16.2
		Bloomz	-	-	-	-	-	3
		Cohere	3	2.3	3	3	2.3	3
		ChatGPT	3	2.3	3	3	3	3
		Davinci	3	2.3	3	3	3	3
		Dolly	3	2.3	3	3	2.7	3
Test		GPT4	-	-	-	-	-	3
			0.5	0.5	0.5	0.5	0.5	-
Dev		Bloomz	0.5	0.5	0.5	0.5	0.5	-
			0.5	0.5	0.5	0.5	0.5	-

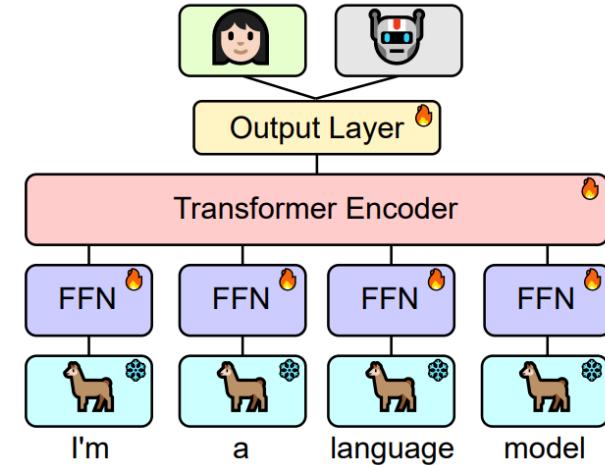
Modeling hypothesis

- LLMs are decoded to sample from regions of high probability, but not the humans
- LLMs assign high probability to constructions with low diversity
- Most of the current LLMs have similar architecture and training data → learn similar text distributions
- Let's extract token-level probabilistic features from a set of LLMs and use them as features:



LLMIXTIC

- Four 8-bit quantized LLaMA-2 models: 7b, 7b-chat, 13b, and 13b-chat
- Three probabilistic features from each LLM, for each token i :
 - Log probability of the predicted token $\alpha_i = \max_{y \in \mathcal{V}} \log p_\theta(y|x_{<i})$
 - Entropy of the distribution $\beta_i = -\sum_{y \in \mathcal{V}} p_\theta(y|x_{<i}) \log p_\theta(y|x_{<i})$
 - Log probability of the observed token $\gamma_i = \log p_\theta(x_i|x_{<i})$
- Each token is represented with 12 features (4 LLMs, 3 features), and a text sequence as $X \in \mathbb{R}^{L \times 12}$
- Each text sequence is passed to an FFN and 1-layer, 4-head Transformer encoder to mix the features.
- Just 0.0002% of the parameters are updated, the remaining ones are frozen (LLaMA models).



Experiments (model comparison)

- We compared 6 models on the provided dev set:
 - LR+READ: logistic regression with readability features
 - LR+TFIDF: logistic regression with TFIDF features
 - Roberta-base: encoder model finetuned for the task
 - E5-base: sentence encoder model finetuned for the task
 - LLMIXTIC (GPT-2): using GPT-2 to extract features
 - LLMIXTIC (LLaMA-2): using LLaMA-2 to extract features
- The scale of the LLM for feature extraction has a large impact
- Also, the similarity w.r.t to the LLMs of the task (e.g., instruction tuning)

Model	Accuracy (%)
LR+READ	42.32
LR+TFIDF	61.26
roberta-base	80.58
e5-base	74.48
LLMIXTIC (w/ GPT-2)	67.42
LLMIXTIC (w/ LLaMA-2)	85.98

Experiments (LLM/feature ablation)

- Including instruction tuned models largely improves the results (by the dataset's nature)
- Just with non-instruction tuned models, LLMIXTIC (LLaMA-2) outperforms LLMIXTIC (GPT-2)
- Comparable with the best neural baseline just using the log prob of the predicted token
- The more features added, the better the results

Ablation	Configuration	Accuracy (%)
LLMs	LLaMA-v2-7b	74.90
	+ LLaMA-v2-13b	75.86
	+ LLaMA-v2-7b-chat	78.48
	+ LLaMA-v2-13b-chat	85.98
Features	Predicted	79.40
	+ Entropy	83.26
	+ Observed	85.98

Test results (official ranking)

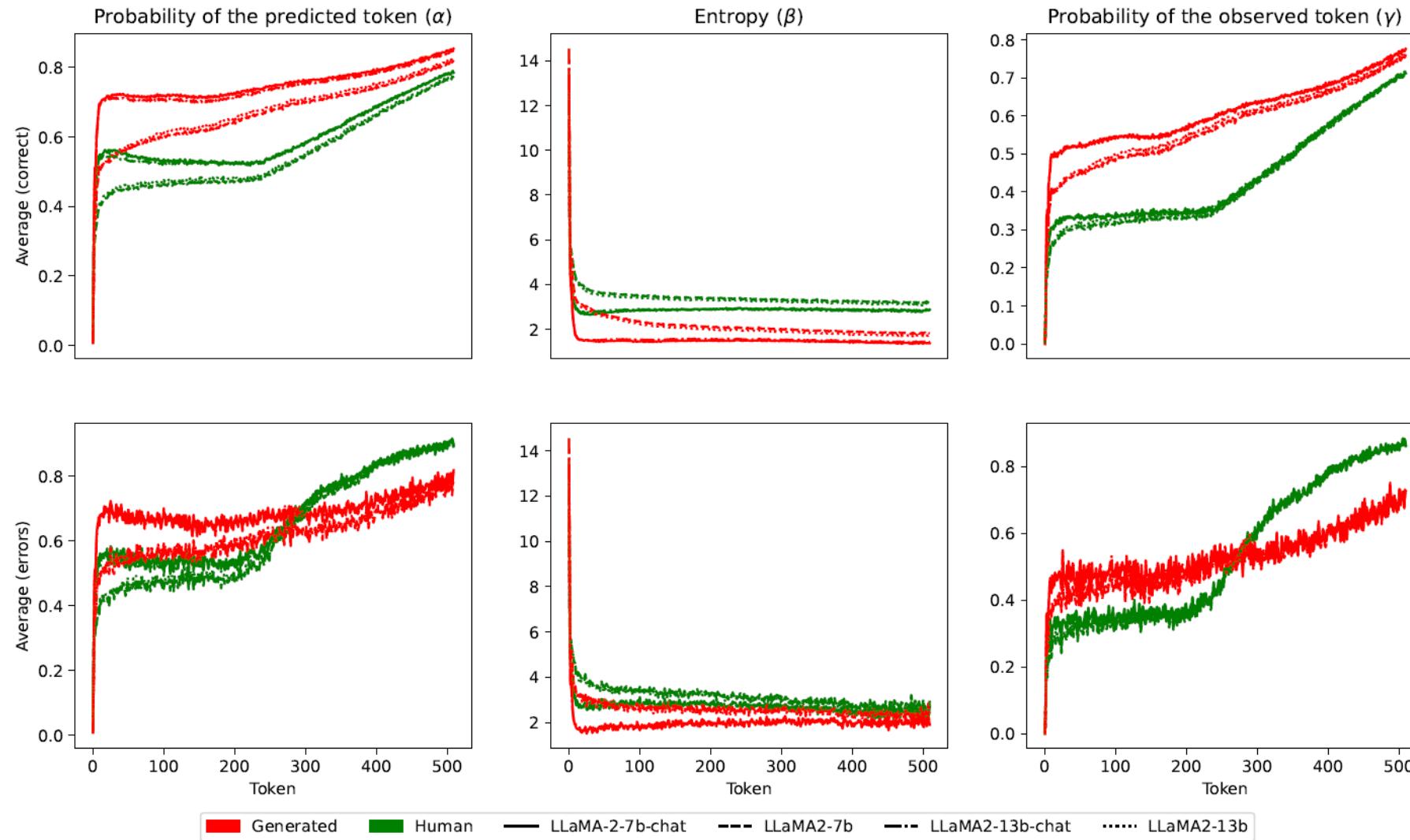
Track	Rank	Name	Accuracy (%)
Monolingual	1	Genaios	96.88
	2	USTC-BUPT	96.09
	20	<i>baseline</i> (119 more)	88.46
Multilingual	1	USTC-BUPT	95.98
	14*	Genaios	89.97
	25	<i>baseline</i> (44 more)	80.88

Analysis

- We examined the probabilistic features, the attention heads, and text patterns in misclassified samples
- We found:
 - LLMIXTIC fails when human text probabilities become larger than for generated texts
 - LLMIXTIC pays more attention to the last positions of the sequences
 - Human text is more often confused with generated text than vice versa (FPR: 4.38%, FNR: 1.97%)
 - Newlines are predominant in false negatives

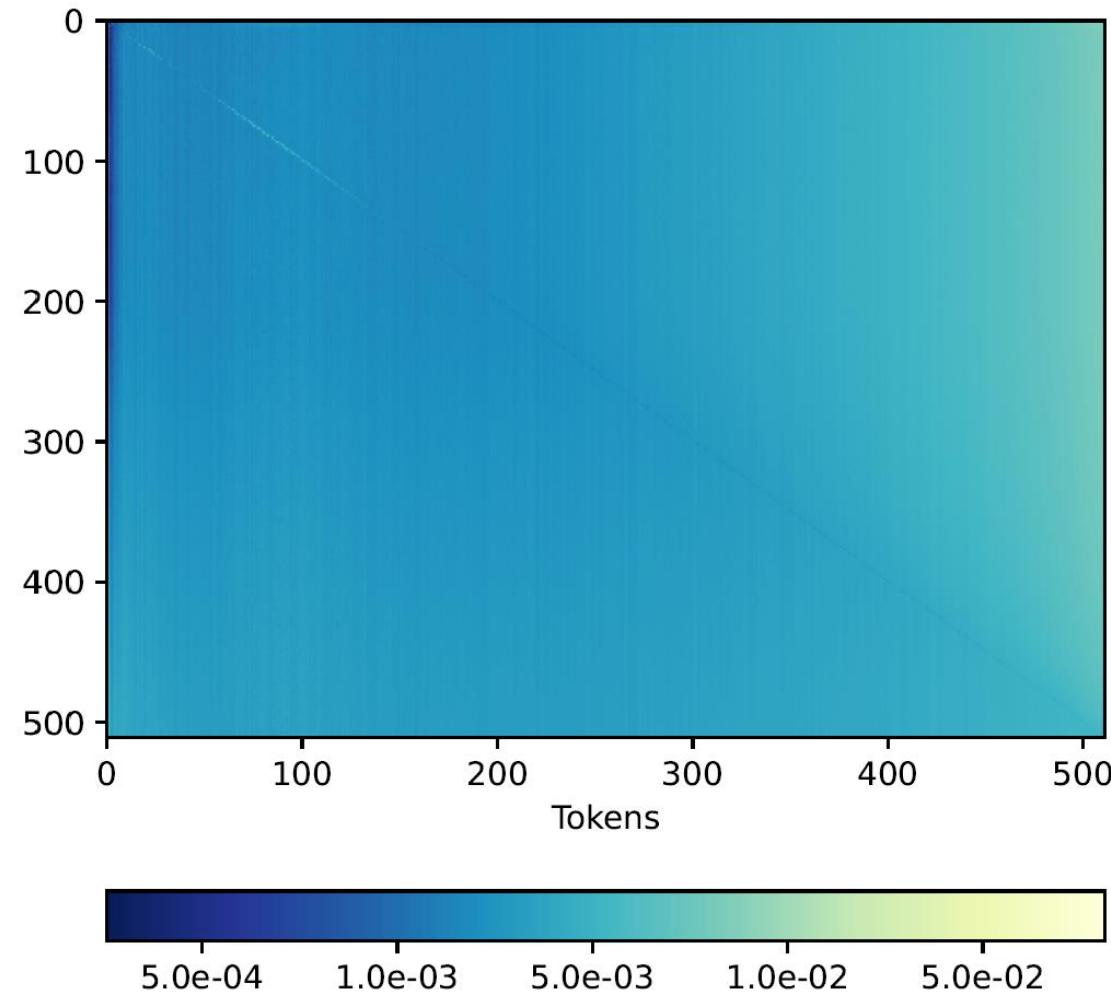
Analysis (probabilistic features)

LLMIXTIC fails when human text probabilities become larger than for generated texts



Analysis (attention scores)

LLMIXTIC pays more attention to the last positions of the sequences



Sample-averaged, head-averaged attention scores from LLMIXTIC's Transformer encoder

Conclusions

- We proposed **LLMIXTIC**, a **Transformer encoder** that mixes **token-level probabilistic features** from LLMs
- Obtains the **best results** in the competition, **20%** of accuracy above the average participant
- LLMIXTIC works **better** when the **human text probabilities** are consistently **lower** than the **generated** ones...
- which is due to the **patterns** in the **attention scores** the encoder is learning

Genaios at SemEval-2024 Task 8: Detecting Machine-Generated Text by Mixing Language Model Probabilistic Features

Areg Mikael Sarvazyan and José Ángel González and Marc Franco-Salvador
Genaios, Valencia, Spain
{areg.sarvazyan, jose.gonzalez, marc.franco}@genaios.ai

Abstract

This paper describes the participation of the Genaios team in the monolingual track of Subtask A at SemEval-2024 Task 8. Our best system, LLMIXTIC, is a Transformer Encoder that mixes token-level probabilistic features extracted from four LLaMA-2 models. We obtained the best results in the official ranking (96.88% accuracy), showing a false positive ratio of 4.38% and a false negative ratio of 1.97% on the test set. We further study LLMIXTIC through ablation, probabilistic, and attention analyses, finding that (i) performance improves as more LLMs and probabilistic features are included, (ii) LLMIXTIC puts most attention on the features of the last tokens, (iii) it fails on samples where human text probabilities become consistently higher than for generated text, and (iv) LLMIXTIC's false negatives exhibit a bias towards text with newlines.

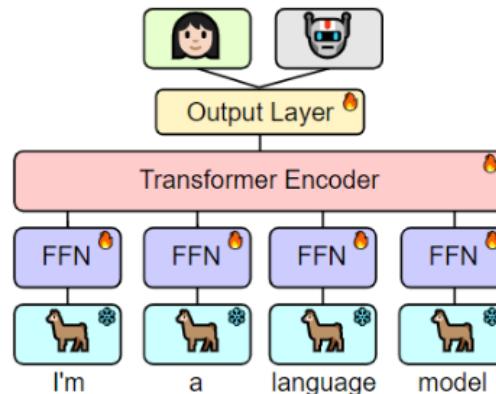


Figure 1: Overview of the proposed system. Modules marked with \spadesuit are frozen. Those with \heartsuit are trainable.

our solution as the Genaios team at SemEval-2024 Task 8: *Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection* (Wang et al., 2024a).

Our starting point is the observation that LLMs

SYMANTO PSYCHOLOGY AI

IberAuTexTification 2024: Automated Text Identification on Languages of the Iberian Peninsula

Areg Mikael Sarvazyan, Jose Ángel González,
Marc Franco-Salvador

¹ Symanto Research Spain



Participate!

- Next iteration of **AuTexTification**:
 - More **languages** (*Spanish, Basque, Catalan, Galician, English, Portuguese*)
 - More **domains**
 - More **LLMs**
- High **quality, unbiased** dataset generated with **TextMachina**
- Participation is open until 3rd June (**deadline for submitting results**):
 - Webpage: <https://sites.google.com/view/iberautextification>
 - Github: <https://github.com/Genaios/IberAuTexTification>
 - Slack channel: https://join.slack.com/t/iberautextification/shared_invite/zt-2gd3aa8wn-cg3nx910t~mErzFXFm1ysw&sa=D&sntz=1&usg=AOvVaw2jXbnWaFAnmvWG7IAox9LN
 - Contact us: organizers.autextification@gmail.com



On behalf of the presenters and myself as an AI, I want to extend our heartfelt gratitude to all of you for your active engagement.

Thanks!

