AuTexTification 2023 and more

Detection and Attribution of Machine-Generated Text in Multiple Domains

Areg Mikael Sarvazyan

areg.sarvazyan@symanto.com

Contents

- The problem: detecting machine-generated text
- Possible solutions
- The AuTexTification 2023 shared task
- ...and more: Generalization to model families and parameter scales

Machine-Generated Text (MGT)

Text that has been produced without human intervention

- Large-scale automatic text generation

- Sampling from a language model
- Before LLMs
  - Low quality
  - Easy to distinguish from human text
  - Factual errors
  - Syntactic and grammar artifacts

Machine-Generated Text (MGT)

- Now we have **Large Language Models!**
  - High-quality multi-domain and multi-style generation
  - Factual errors [1], *hallucination*
  - **Can be used to generate high-quality malicious text very easily**
- **Ensure a responsible use of LLMs**
  - Detect machine-generated text
  - Attribute machine-generated text to a particular model
    - Who is behind malicious MGT?
    - Important for fair use and licensing

How to detect MGT?

- **Zero-shot [2, 3]**
  - Usually white-box
  - Use *model A* probabilities to detect *model A* text
  - Vast LLM ecosystem
    - Not generalizable to detecting MGT from other models

How to detect MGT?

- **Supervised [4-8]**
  - Train models on annotated text and its linguistic and statistical features
  - Could generalize to other text generation models
  - **Need high quality multi-domain/style data**

AuTexTification 2023

AuTexTification 2023

Shared task @IberLEF2023

- **Annotated multi-domain data in Spanish and English**
- Study generalization to new domains
- Tasks: **MGT Detection** and **MGT Attribution**

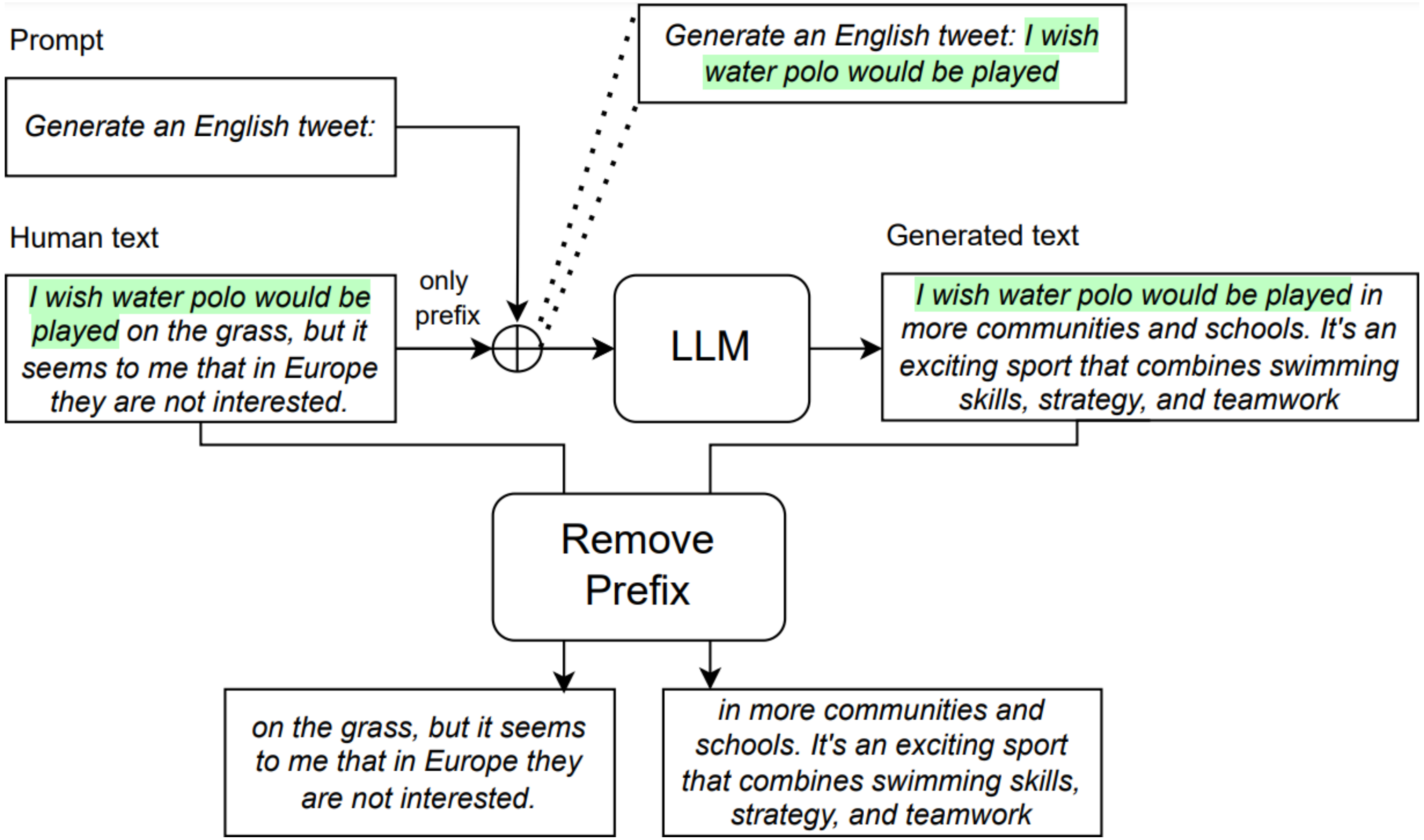
AuTexTification 2023

MGT Detection

- human or machine

MGT Attribution

- Which model's MGT?



AuTexTification 2023

Final datasets

- Generations by BLOOM and GPT models with nucleus sampling
  - BLOOM-1b1, -3b, -7b
  - GPT: babbage (1b), curie (6b7), davinci (175b)
- Domains: **tweets, reviews, how-to articles, news, legal documents**
- Base datasets with balanced domains:
  - English: Amazon Polarity [9], XSUM [10], WikiLingua [11], MultiEURLEX [12], TSATC [13]
  - Spanish: COAH [14], COAR [15], MLSUM [16], XLSum [17], WikiLingua [11], MultiEURLEX [12], Spanish Politics Tweets [18]
- **Human continuations** and **generated continuations**
- Cleaning punctuations, whitespaces & filtering by language ID, empty generations, etc.

AuTexTification 2023

Baselines

- Fine-tuned transformers: Roberta-BNE [19] (Spanish), DeBERTaV3 [20] (English)
- Symanto Brain[1]: Zero and few-shot models (SB)
- Random baseline

Submissions included

- Token and text level probabilities and entropies
- Lexical, syntactical, grammatical and readability text features
- Text embeddings: CNNs, pre-trained transformers
- Logistic Regression, MLPs, Tree-based classifiers, fine-tuned transformers
- **Best results are ensembles of many classifiers on many feature combinations**

1 PEFT fine-tuning and classification by embedding similarities to label descriptions. More info: https://www.symanto.com/nlp-tools/symanto-brain/.

AuTexTification 2023

Subtask 1: Machine-Generated Text Detection

- Train: Tweets, how-to articles, legal documents
- Test: Reviews, news

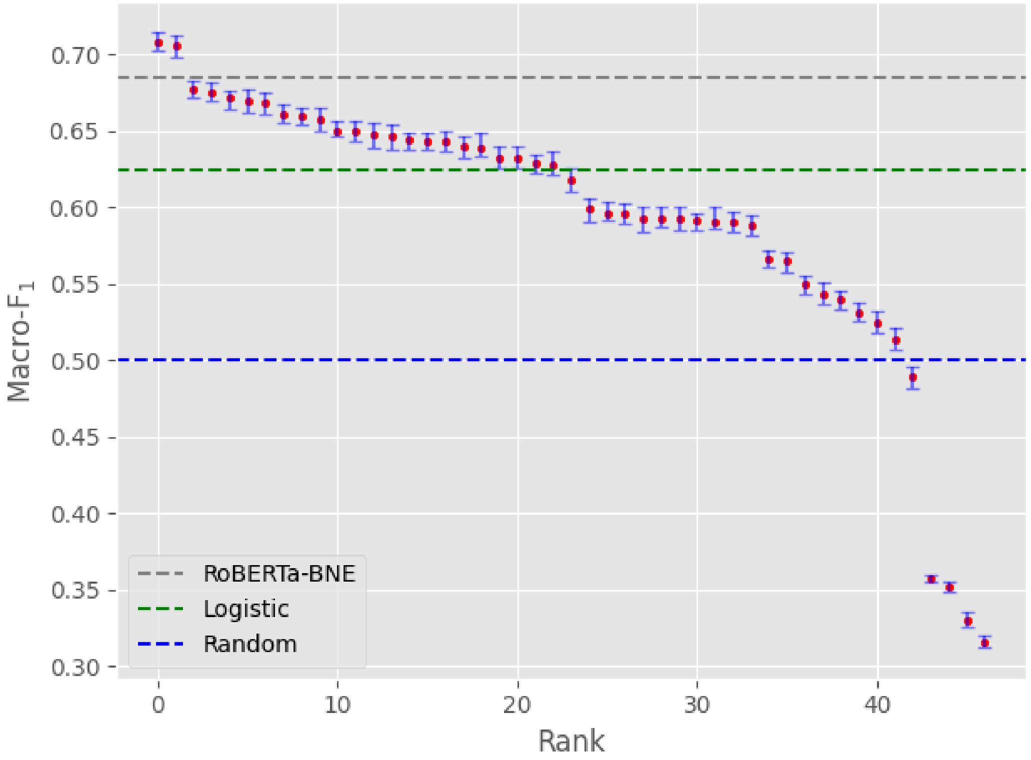
	Language	Split	**Human**	**Generated**
Subtask 1: MGT Detection	English	Train	17,046	16,799
		Test	10,642	11,190
	Spanish	Train	15,787	16,275
		Test	11,209	8,920

AuTexTification 2023

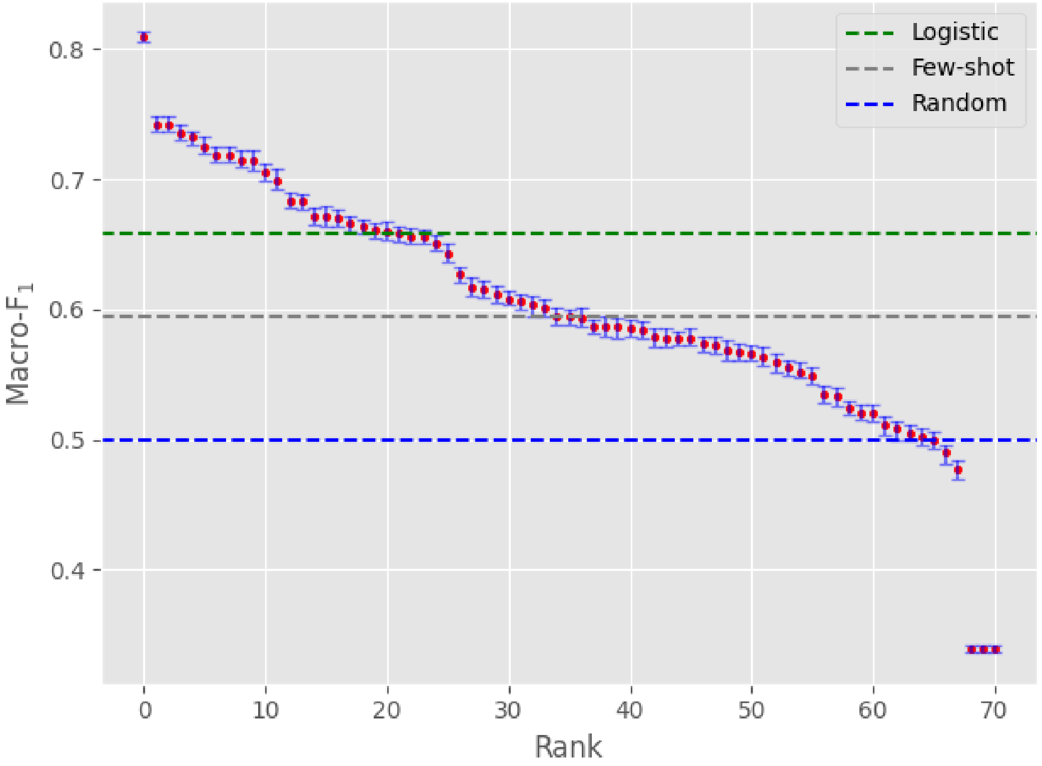
Subtask 1: Machine-Generated Text Detection

- Rank and macro-f1 w/bootstrapped confidence intervals

Spanish



English

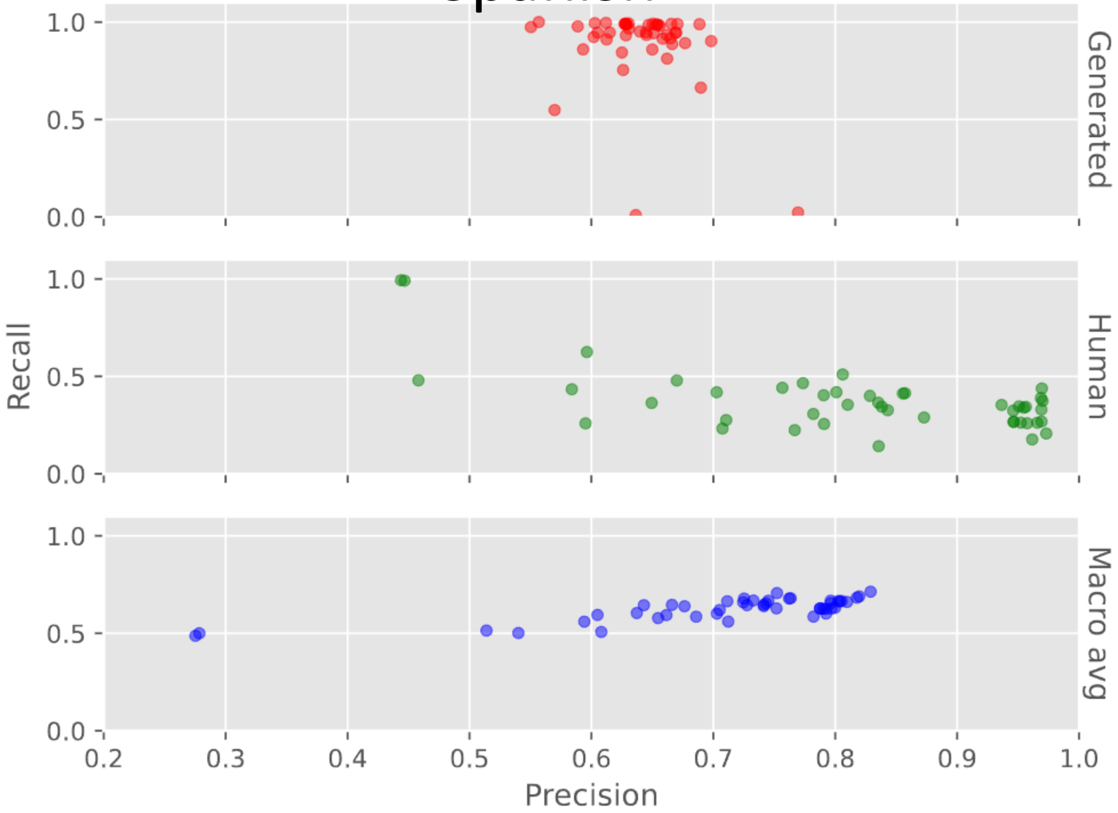


AuTexTification 2023

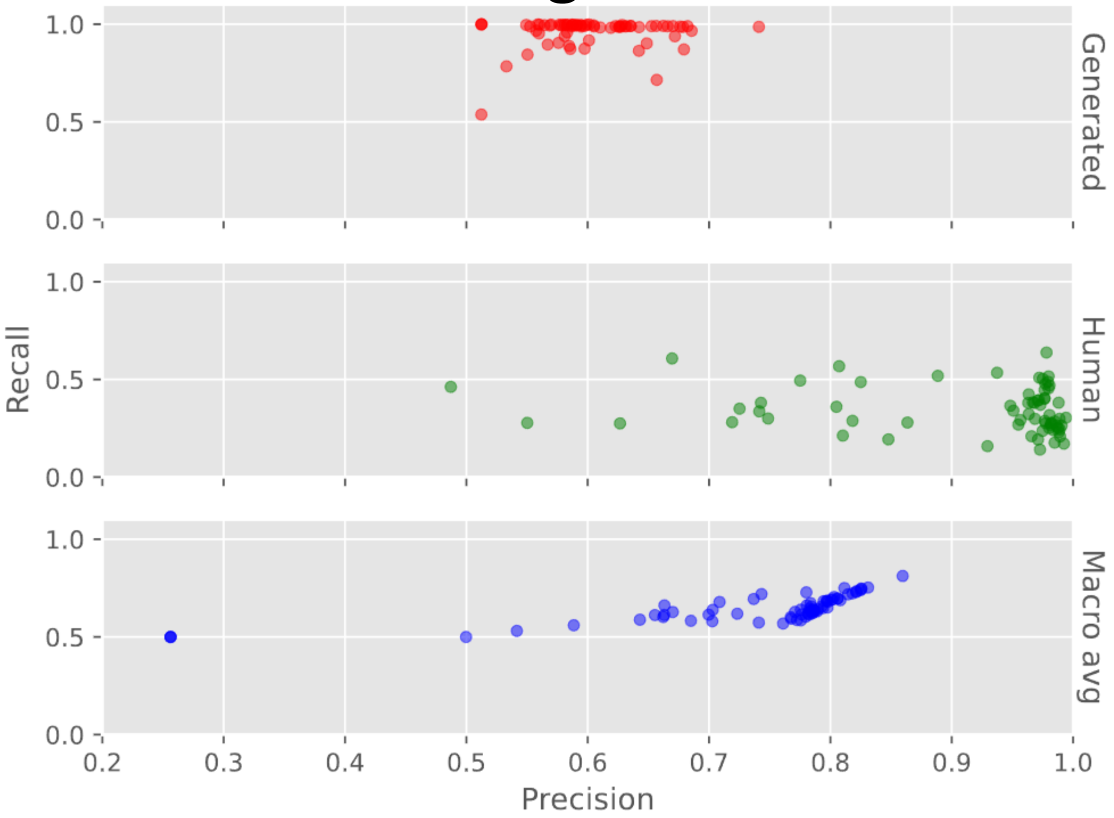
Subtask 1: Machine-Generated Text Detection

- Precision-recall curves

Spanish



English

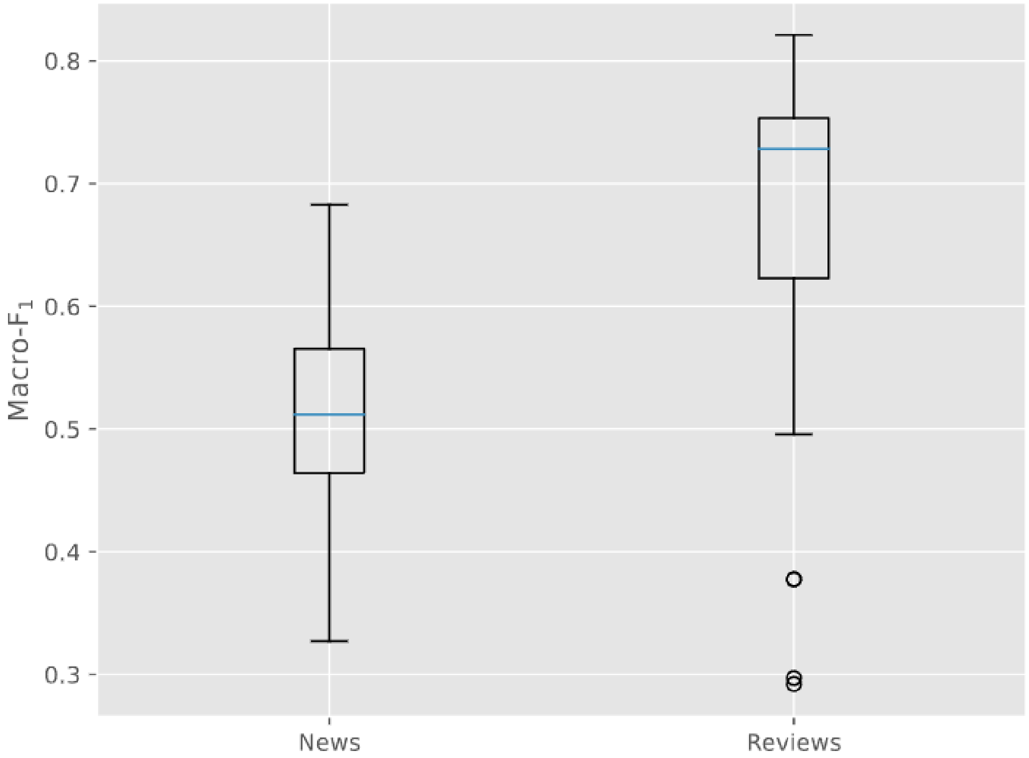


AuTexTification 2023

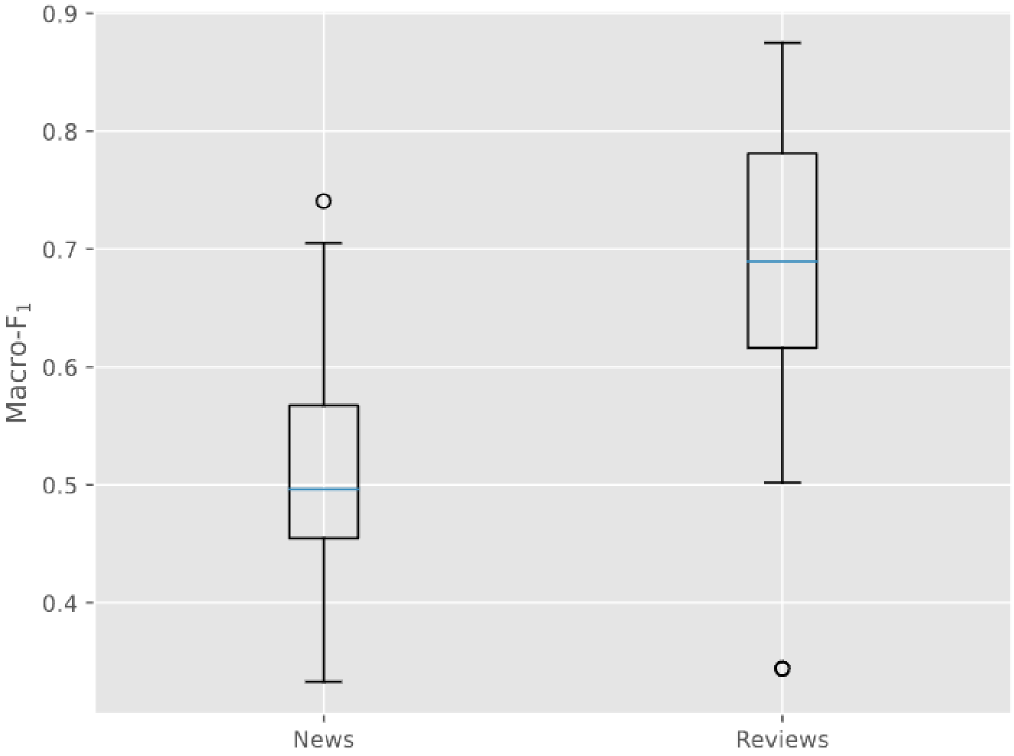
Subtask 1: Machine-Generated Text Detection

- Per domain macro-F1

Spanish



English



AuTexTification 2023

Subtask 2: Model Attribution

- Same domains for train and test (tweets, how-to, reviews, news, legal)

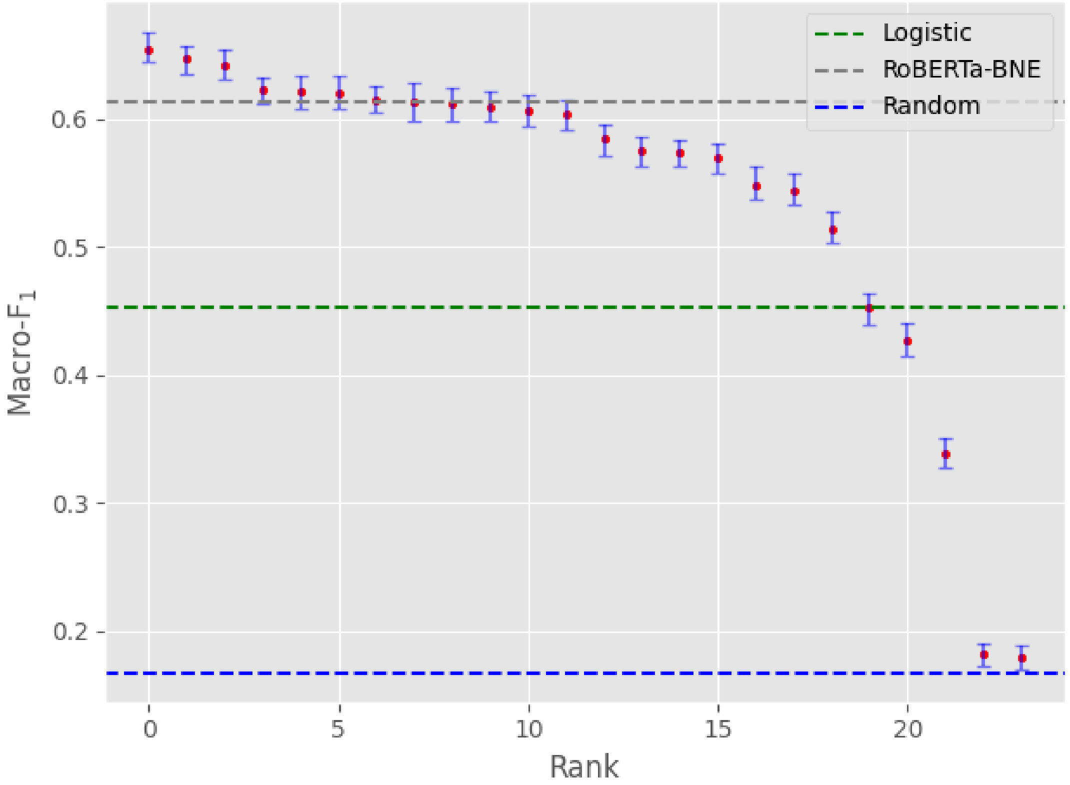
			BLOOM			GPT		
	Language	Split	1b7	3b	7b	babbage	curie	davinci
Subtask 2: Model Attribution	English	Train	3,562	3,648	3,687	3,870	3,822	3,827
		Test	887	875	952	924	979	988
	Spanish	Train	3,422	3,514	3,575	3,788	3,770	3,866
		Test	870	867	878	946	1,004	917

AuTexTification 2023

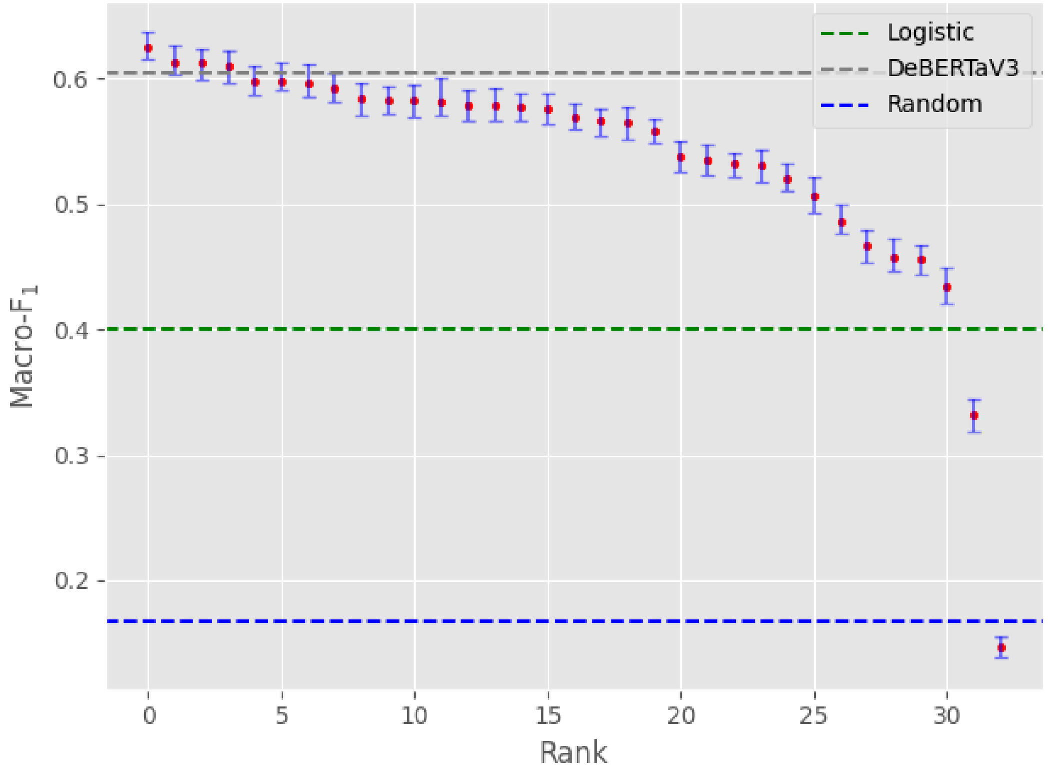
Subtask 2: Model Attribution

- Rank and macro-f1 w/bootstrapped confidence intervals

Spanish



English

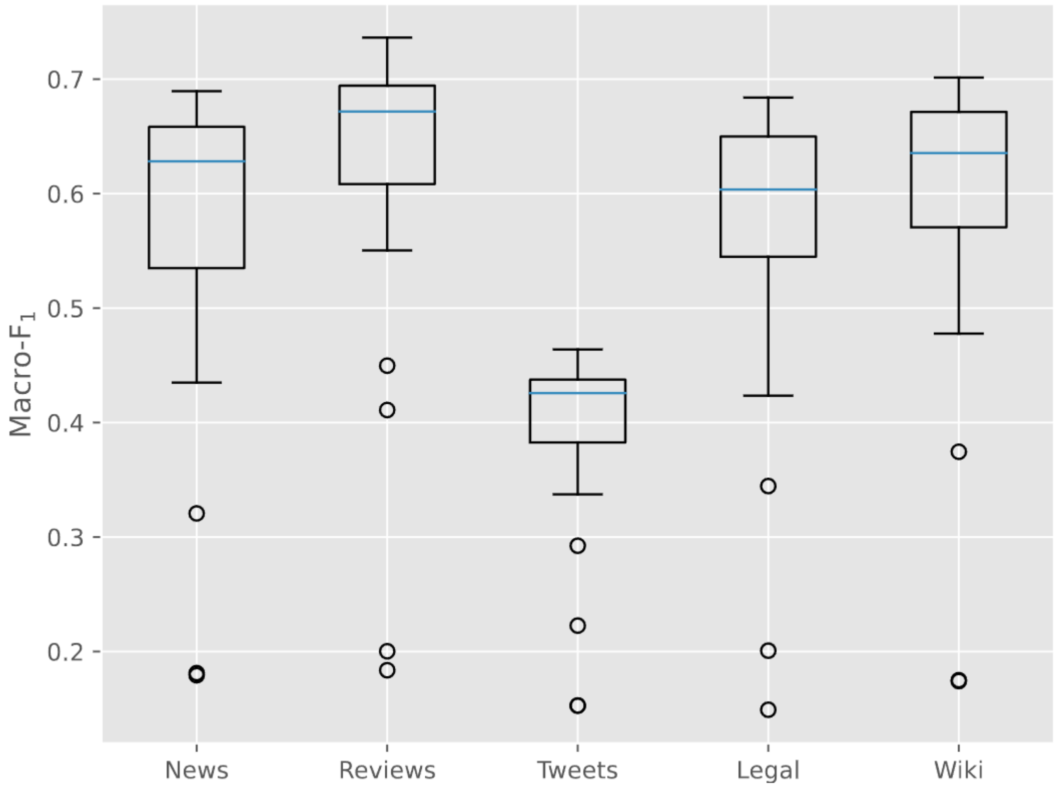


AuTexTification 2023

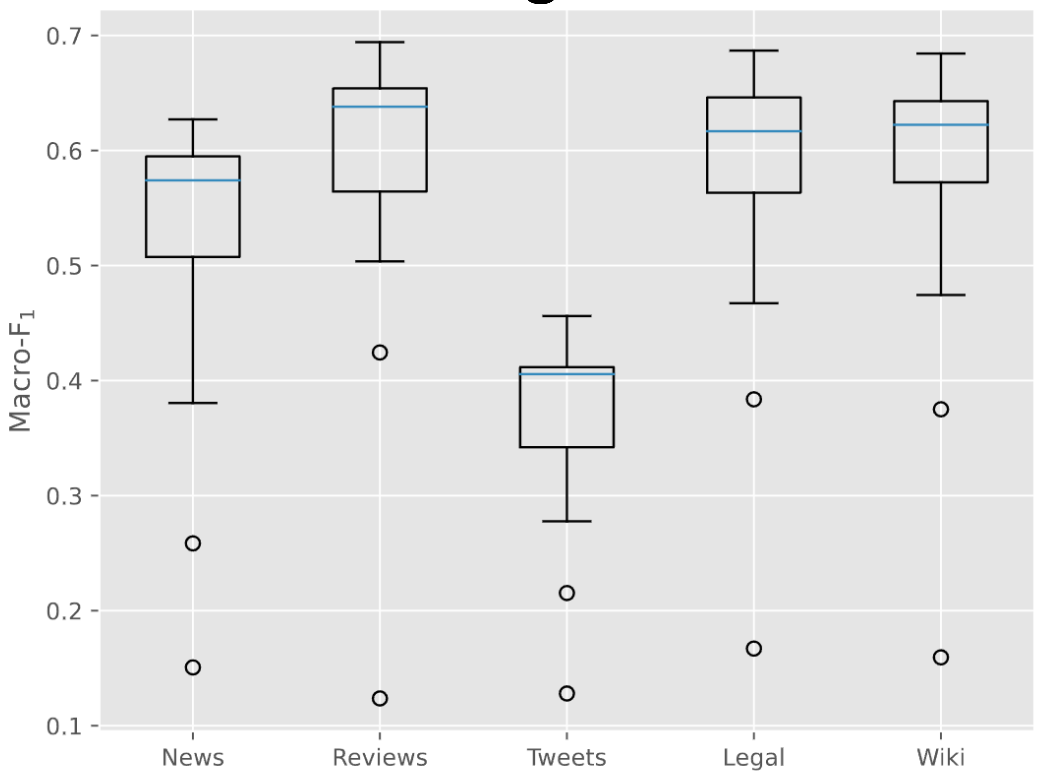
Subtask 2: Model Attribution

- Per domain macro-F1

Spanish



English



Generalization to Families and Scales

Generalization to Families and Scales

Study MGT detector generalization

- *Family:* models trained similarly (BLOOM is one family, GPT is another)
- *Scale:* models of similar number of parameters

- Fine-tune pre-trained transformers on one family (scale)
- Evaluate on other family (scale)

Generalization to Families and Scales

Study MGT detector generalization

- Group data from both subtasks:
  - Human text from subtask 1
  - Generated text with fine-grained annotations from subtask 2
- Data transformations for model, class and domain balance
  - All five domains in both train and test sets
- Train and test splits for each family or scale
- 3 MGT detectors: fine-tuned BLOOM-560m, DeBERTaV3, XLM-RoBERTa
- We only present results for English

Generalization to Families and Scales

Generalization to families

		BLOOM			GPT		
Train	Detector	Gen	Hum	Mean	Gen	Hum	Mean
BLOOM	BLOOM-560	93.70	93.92	93.81	59.32	75.81	67.57
	DeBERTa	**95.21**	**94.79**	**95.00**	76.19	80.66	78.43
	XLM-R	93.13	92.14	92.63	**79.26**	**80.86**	**80.06**
GPT	BLOOM-560	72.17	79.82	75.99	89.61	89.78	89.69
	DeBERTa	**85.61**	**85.05**	**85.33**	**89.94**	**87.82**	**88.88**
	XLM-R	82.40	82.04	82.22	89.52	87.22	88.37

Generalization to Families and Scales

Generalization to families

		BLOOM			GPT		
Train	Detector	Gen	Hum	Mean	Gen	Hum	Mean
BLOOM	BLOOM-560	93.70	93.92	93.81	59.32	75.81	67.57
	DeBERTa	**95.21**	**94.79**	**95.00**	76.19	80.66	78.43
	XLM-R	93.13	92.14	92.63	**79.26**	**80.86**	**80.06**
GPT	BLOOM-560	72.17	79.82	75.99	89.61	89.78	89.69
	DeBERTa	**85.61**	**85.05**	**85.33**	**89.94**	**87.82**	**88.88**
	XLM-R	82.40	82.04	82.22	89.52	87.22	88.37

Generalization to Families and Scales

Generalization to families

		BLOOM			GPT		
Train	Detector	Gen	Hum	Mean	Gen	Hum	Mean
BLOOM	BLOOM-560	93.70	93.92	93.81	59.32	75.81	67.57
	DeBERTa	**95.21**	**94.79**	**95.00**	76.19	80.66	78.43
	XLM-R	93.13	92.14	92.63	**79.26**	**80.86**	**80.06**
GPT	BLOOM-560	72.17	79.82	75.99	89.61	89.78	89.69
	DeBERTa	**85.61**	**85.05**	**85.33**	**89.94**	**87.82**	**88.88**
	XLM-R	82.40	82.04	82.22	89.52	87.22	88.37

Generalization to Families and Scales

Generalization to scales

		1b			7b			175b		
Train	Detector	Gen	Hum	Mean	Gen	Hum	Mean	Gen	Hum	Mean
1b	BLOOM-560	89.69	90.04	89.89	85.22	86.45	85.84	76.37	83.43	79.90
	DeBERTa	**93.46**	**92.88**	**93.17**	**91.84**	**91.04**	**91.44**	89.90	**91.45**	90.67
	XLM-R	89.29	86.96	88.13	87.87	84.67	86.27	**91.12**	90.86	**90.99**
7b	BLOOM-560	87.49	**88.25**	**87.87**	86.02	**86.72**	**86.37**	79.16	84.75	81.96
	DeBERTa	**88.71**	85.99	87.35	**87.20**	83.14	85.17	**92.38**	**92.03**	**92.20**
	XLM-R	86.92	82.89	84.91	85.30	79.59	82.45	90.02	88.87	89.44
175b	BLOOM-560	56.14	74.47	65.30	64.47	77.36	70.92	91.52	**91.97**	91.75
	DeBERTa	69.77	75.51	72.64	**81.36**	**81.86**	**81.61**	**92.64**	91.48	**92.06**
	XLM-R	**73.31**	**75.67**	**74.49**	**81.36**	80.61	80.99	90.50	88.45	89.48

Generalization to Families and Scales

Generalization to scales

		1b			7b			175b		
Train	Detector	Gen	Hum	Mean	Gen	Hum	Mean	Gen	Hum	Mean
1b	BLOOM-560	89.69	90.04	89.89	85.22	86.45	85.84	76.37	83.43	79.90
	DeBERTa	**93.46**	**92.88**	**93.17**	**91.84**	**91.04**	**91.44**	89.90	**91.45**	90.67
	XLM-R	89.29	86.96	88.13	87.87	84.67	86.27	**91.12**	90.86	**90.99**
7b	BLOOM-560	87.49	**88.25**	**87.87**	86.02	**86.72**	**86.37**	79.16	84.75	81.96
	DeBERTa	**88.71**	85.99	87.35	**87.20**	83.14	85.17	**92.38**	**92.03**	**92.20**
	XLM-R	86.92	82.89	84.91	85.30	79.59	82.45	90.02	88.87	89.44
175b	BLOOM-560	56.14	74.47	65.30	64.47	77.36	70.92	91.52	**91.97**	91.75
	DeBERTa	69.77	75.51	72.64	**81.36**	**81.86**	**81.61**	**92.64**	91.48	**92.06**
	XLM-R	**73.31**	**75.67**	**74.49**	**81.36**	80.61	80.99	90.50	88.45	89.48

Generalization to Families and Scales

Generalization to scales

		1b			7b			175b		
Train	Detector	Gen	Hum	Mean	Gen	Hum	Mean	Gen	Hum	Mean
1b	BLOOM-560	89.69	90.04	89.89	85.22	86.45	85.84	76.37	83.43	79.90
	DeBERTa	**93.46**	**92.88**	**93.17**	**91.84**	**91.04**	**91.44**	89.90	**91.45**	90.67
	XLM-R	89.29	86.96	88.13	87.87	84.67	86.27	**91.12**	90.86	**90.99**
7b	BLOOM-560	87.49	**88.25**	**87.87**	86.02	**86.72**	**86.37**	79.16	84.75	81.96
	DeBERTa	**88.71**	85.99	87.35	**87.20**	83.14	85.17	**92.38**	**92.03**	**92.20**
	XLM-R	86.92	82.89	84.91	85.30	79.59	82.45	90.02	88.87	89.44
175b	BLOOM-560	56.14	74.47	65.30	64.47	77.36	70.92	91.52	**91.97**	91.75
	DeBERTa	69.77	75.51	72.64	**81.36**	**81.86**	**81.61**	**92.64**	91.48	**92.06**
	XLM-R	**73.31**	**75.67**	**74.49**	**81.36**	80.61	80.99	90.50	88.45	89.48

Conclusions

- AuTexTification
  - Multi-domain / style annotated datasets for MGT detection and attribution
  - Many types of solutions
  - Scores as high as 80% macro-f1 (detection) and 65% (attribution)
- Family and scale generalization
  - Usually generalize well to families and scales
  - Difficult to generalize when gpt-3 davinci (175B) is involved
    - Quality differences between generated texts subjectively
    - Not so much between human and generated

Questions?

References

[1] Tam, D., Mascarenhas, A., Zhang, S., Kwan, S., Bansal, M., & Raffel, C. (2022). Evaluating the Factual Consistency of Large Language Models Through Summarization. *arXiv preprint arXiv:2211.08412*.

[2] Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., ... & Wang, J. (2019). Release Strategies and the Social Impacts of Language Models.

[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*.

[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] Uchendu, A., Le, T., Shu, K., & Lee, D. (2020, November). Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 8384-8395).

[6] Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020, July). Automatic Detection of Generated Text is Easiest when Humans are Fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1808-1822).

[7] 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).

[8] 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*.

[9] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, *28*.

[10] Narayan, S., Cohen, S. B., & Lapata, M. (2018). Don’t Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1797-1807).

References

[11] Ladhak, F., Durmus, E., Cardie, C., & Mckeown, K. (2020, November). WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4034-4048).

[12] Chalkidis, I., Fergadiotis, M., & Androutsopoulos, I. (2021, November). 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* (pp. 6974-6996).

[13] Naji, I. (2012). TSATC: Twitter Sentiment Analysis Training Corpus.

[14] Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M. T., & Urena-López, L. A. (2014). Cross-domain sentiment analysis using Spanish opinionated words. In *Natural Language Processing and Information Systems: 19th International Conference on Applications of Natural Language to Information Systems, NLDB 2014, Montpellier, France, June 18-20, 2014. Proceedings 19* (pp. 214-219). Springer International Publishing.

[15] Molina-González, M. D., Martínez-Cámara, E. COAR. https://sinai.ujaen.es/en/research/resources/coar.

[16] Scialom, T., Dray, P. A., Lamprier, S., Piwowarski, B., & Staiano, J. (2020, November). MLSUM: The Multilingual Summarization Corpus. In *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 8051-8067). Association for Computational Linguistics.

[17] Hasan, T., Bhattacharjee, A., Islam, M. S., Mubasshir, K., Li, Y. F., Kang, Y. B., ... & Shahriyar, R. (2021, August). XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 4693-4703).

[18] Moya, R. Tweets Política España, Version 6. Retrieved February 1, 2023 from https://www.kaggle.com/datasets/ricardomoya/tweets-poltica-espaa.

[19] Gutiérrez Fandiño, A., Armengol Estapé, J., Pàmies, M., Llop Palao, J., Silveira Ocampo, J., Pio Carrino, C., ... & Villegas, M. (2022). MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, *68*.

[20] He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Este proyecto ha sido cofinanciado por el
Fondo Europeo de Desarrollo Regional (FEDER)
con el objetivo de promover el desarrollo tecnológico,
la innovación y una investigación de calidad.

Una manera de hacer Europa

SYMANTO SPAIN, S.L.U.
PRO²HATERS: “PROactive PROfiling HATE speech spreadeRS”

UNIÓN EUROPEA

MINISTERIO DE CIENCIA E INNOVACIÓN

CDTI

XAI-DisInfodemics:
eXplainable AI for disinformation and conspiracy detection during infodemics

MINISTERIO DE CIENCIA E INNOVACIÓN

Financiado por la Unión Europea
NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

AGENCIA ESTATAL DE INVESTIGACIÓN

Grant **PLEC2021-007681**
funded by MCIN/AEI/ 10.13039/501100011033
and by European Union NextGenerationEU/PRTR.

ANDHI:

ANomalous Difussion of Harmful Information

MINISTERIO DE CIENCIA E INNOVACIÓN

Financiado por la Unión Europea NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

AGENCIA ESTATAL DE INVESTIGACIÓN

Grant **CPP2021-008994**

funded by MCIN/AEI/ 10.13039/501100011033

and by European Union NextGenerationEU/PRTR.