Supervised Machine-Generated Text Detectors: Family and Scale Matters

18th September, 2023

Areg Mikael Sarvazyan

areg.sarvazyan@symanto.com

José Ángel González

Marc Franco-Salvador

Paolo Rosso

Contents

Introduction

Motivation

- **AI Democratization + LLMs: Possible to generate high-quality malicious text very easily**
- Content moderation and defense against large-scale malicious MGT (spam, propaganda, …)
- Ensure AI regulations and licenses are followed
- Maintain high-quality text data for future training of language models
- **Ensure responsible usage of LLMs**
  - Detecting MGT (binary classification)
  - Attributing MGT to a particular model (N-way classification)

Machine-Generated Text

Text that has been produced without human intervention

- Generated with LLMs
  - High-quality multi-domain and multi-style generation
  - Factual errors [1], *hallucination*
  - AI democratization = everyone has access
    - **Anyone can generate malicious texts**

		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

We focus on **large-scale** and **high-accessibility**.

[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*.

Machine-Generated Text

State of the Art

- **Waterkmarking** [2]**:** make MGT self-identifiable through cryptographic watermarks
    - Only possible if everyone enforces watermarks (otherwise: can paraphrase with another model)
- **Machine-aided** [3]**:** capture text artifacts automatically to help humans detect MGT
- **Zero-shot** [4] (white-box)
    - Use a LLMs probabilities to detect its own MGT: **not generalizable to new model**
        - **We usually don't know what models generated the texts**
        - **Could not have white-box access to it**

[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] 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*.

[4] 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*

Machine-Generated Text

State of the Art

- **Supervised** [5, 6]
  - Train models on text and its linguistic and statistical features: **generalization is possible**
  - Need high quality multi-domain/style data
  - Transformer-based models studied under single-domain assumption
  - Generalization capabilities to new domains must be studied [7]

- **MGT attribution is an open problem** [8]
  - Only one work studied it deeply with simple models [9]

[5] 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).
[6] 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).
[7] 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.
[8] Crothers, E., Japkowicz, N., & Viktor, H. L. (2023). Machine-generated Text: A Comprehensive Survey of Threat Models and Detection Methods. *IEEE Access*.
[9] 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).

Machine-Generated Text

In this work

- **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?

Definitions

- **Family:** group of models trained in the same manner
- **Scale:** group of models with similar number of parameters

Generalization of MGT Detectors

Dataset

- AuTexTification 2023 [10] dataset
- Balanced by class, domain, text generation model, language
  - Subtask 1: MGT detection
  - Subtask 2: 6-way Attribution
- MGT by BLOOM and GPT
- 5 domains, 2 languages

		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

[10] 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.

Generalization of MGT Detectors to new Families and Scales

Generalization of MGT Detectors

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

Datasets

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

To Unseen Model Families

- Great results when not generalizing to new families

		BLOOM			GPT		
Train	Detector	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

To 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

		BLOOM			GPT		
Train	Detector	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

To Unseen Model Families

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

		BLOOM			GPT		
Train	Detector	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

To Unseen Parameter Scales

- Great performance when not generalizing to unseen scales

		1b			**7b**			**175b**		
Train	Detector	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

To Unseen Parameter Scales

- Great performance in some generalization scenarios

		1b			7b			175b		
Train	Detector	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

To Unseen Parameter Scales

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

		1b			7b			175b		
Train	Detector	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

To Unseen Parameter Scales

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

		1b			7b			175b		
Train	Detector	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 preferible
  - **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

Datasets

	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)

Attribution to Families and Scales

Attributing to Families

- Very feasible and practical

	English			Spanish			
Attributor	**BLOOM**	**GPT**	Mean	Attributor	**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**

Attribution to Families and Scales

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

Conclusions and Future Work

Conclusions

- 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

Future Work

- 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”? And how can we use this to our advantage?

Questions?

MGT Detector generalization in Spanish

To Unseen Model Families

		BLOOM			GPT		
Train	Detector	GEN	HUM	Mean	GEN	HUM	mean
BLOOM	BLOOM-560	88.05	87.78	87.91	65.03	73.52	69.28
	MarIA	**96.25**	**96.29**	**96.27**	58.95	75.91	67.43
	XLM-R	91.74	90.32	91.03	**73.93**	**76.29**	**75.11**
GPT	BLOOM-560	52.68	73.91	63.30	90.69	91.12	90.91
	MarIA	56.91	75.64	66.27	**94.97**	**94.98**	**94.98**
	XLM-R	**70.58**	**76.76**	**73.67**	91.14	89.50	90.32

MGT Detector generalization in Spanish

To Unseen Parameter Scales

		1b			7b			175b		
Train	Detector	Gen	Hum	Mean	Gen	Hum	Mean	Gen	Hum	Mean
1b	BLOOM-560	90.57	90.09	90.33	86.76	86.77	86.72	86.58	88.98	87.78
	MarIA	**94.13**	**94.25**	**94.19**	**90.90**	**91.54**	**91.22**	83.33	87.50	85.42
	XLM-R	87.85	84.35	86.10	86.67	82.62	84.64	**91.58**	**91.18**	**91.38**
7b	BLOOM-560	88.03	88.35	88.19	87.54	87.75	87.65	88.48	90.41	89.44
	MarIA	**91.75**	**92.00**	**91.88**	**92.52**	**92.54**	**92.53**	**93.43**	**94.20**	**93.82**
	XLM-R	85.61	80.24	82.92	84.64	78.37	81.51	90.16	88.69	89.43
175b	BLOOM-560	51.85	73.16	62.50	55.37	74.22	64.80	93.27	93.64	93.45
	MarIA	53.77	74.23	64.00	64.16	77.27	70.71	**96.29**	**96.30**	**96.29**
	XLM-R	**73.45**	**75.17**	**74.31**	**79.97**	**78.88**	**79.42**	90.74	88.80	89.77

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.