2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Detecting Automated Generated Text with LLMs

Fernando Aguilar-Canto
Computational Cognitive Sciences Laboratory-CIC
Instituto Politécnico Nacional
Mexico City, Mexico
pherjev@gmail.com

Marco Antonio Cardoso-Moreno
Computational Cognitive Sciences Laboratory-CIC
Instituto Politécnico Nacional
Mexico City, Mexico
mcardosom2021@cic.ipn.mx

Diana Jiménez
Computational Cognitive Sciences Laboratory-CIC
Instituto Politécnico Nacional
Mexico City, Mexico
dianaljl.99@gmail.com

Hiram Calvo
Computational Cognitive Sciences Laboratory-CIC
Instituto Politécnico Nacional
Mexico City, Mexico
hcalvo@cic.ipn.mx

***Abstract*—The development of Large Language Models (LLMs) like GPT-series and BLOOM has revolutionized Artificial Intelligence, yet it has also brought forth challenges in misuse, such as fake content generation and academic cheating. Detecting whether a text is generated by an LLM or written by a human has become imperative. Fine-tuned LLMs have proven to be a promising approach in this regard. In our study, we fine-tuned seven LLMs (BERT, DeBERTa-v3, RoBERTa, XLM-RoBERTa, GPT-2 Medium, GPT-2 Large, GPT-2 XL) to detect text generated by even larger models (GPT-3 and BLOOM) in the AuTexTification task. Among the models, GPT-2 Medium exhibited the best performance in the testing set, achieving an F1-macro score of 0.83272 and an accuracy of 0.83442, surpassing the benchmark's best-known result.**

***Index Terms*—Generated Text Detection, Large Language Models, AuTexTification, GPT-2**

I. Introduction

The rapid advancements in Large Language Models (LLMs) like the GPT series [1]–[4], Pathways Language Model (PaLM) [5], LLaMA [6], and BLOOM [7], have sparked concerns regarding potential misuse [3], [8]–[12]. To address these challenges, a significant task has emerged – detecting Machine Generated Text (MGT), a binary classification problem.

This paper focuses on fine-tuning seven LLMs using the AuTexTification dataset [13], specifically subtask 1 (MGT detection) in English. The results are presented in Section IV, followed by a summary of our findings in Section V.

II. Related work

Large Language Models (LLMs) have shown effectiveness in detecting content generated by other LLMs. For instance, Uchendu *et al.* [9] employed various models like GROVER, GPT-2, GLTR, BERT, and RoBERTa to detect text produced by GPT-3. Similarly, current generative models like ChatGPT and LLaMA have been addressed using LLMs such as RoBERTa [14]–[17], DistilBERT [18], LongFormer [19], and OPT-125M [16].

Regarding the AuTextification task, the top-performing model achieved a F1-macro score of 80.91 (confidence interval: 80.4 to 81.38) [13].

III. Methodology

To tackle this problem, we propose fine-tuning the following LLMs: (1) BERT-cased Large [20], (2) RoBERTa Large [21], (3) XLM-RoBERTa Large [22], (4) DeBERTA Large [23], (5) GPT-2 Medium, (6) Large, and (7) XL [2].

In all cases, default hyperparameters from HuggingFace were used. For GPT-2 re-training, the model served as a feature extractor with a classification layer added, and only one epoch was employed. For BERT-based models, the entire model was fine-tuned using a low learning rate (3e-5) and three epochs. A train-validation split of 80-20 was utilized, and evaluation was performed with the proposed testing set of the task.

IV. Results

The primary outcomes for the BERT-based models are displayed in tables I (validation) and III (testing), while the GPT-2 results are presented in tables II (validation) and IV (testing). Evidently, the most promising comparative results were achieved with GPT-2 models, particularly GPT-2 Medium.

TABLE I
Main results on the validation set of the subtask 1 with the BERT-based models.

Model	1	2	3	4
F1-macro	0.89187	0.90477	0.90647	0.90647
F1-weighted	0.89166	0.90459	0.90633	0.93589
Accuracy	0.89216	0.90501	0.90663	0.93588
Precision	0.82763	0.82763	0.84626	0.90505
Recall	0.90505	0.90505	0.90505	0.90505

V. Conclusions

This paper conducted a comprehensive comparison of five distinct Large Language Models (LLMs) with seven configurations to address the task of detecting text generated by other LLMs (specifically GPT-3 and BLOOM, [13]). Although there were no significant disparities between BERT-based and GPT-2 models in the validation set, notable differences emerged in the testing set. Notably, GPT-2 Medium exhibited superior

TABLE II
MAIN RESULTS ON THE VALIDATION SET OF THE SUBTASK 1 WITH THE GPT-BASED MODELS.

Model	5	6	7
F1-macro	0.88900	0.92701	0.93169
F1-weighted	0.88933	0.92704	0.92704
Accuracy	0.89040	0.92703	0.93176
Precision	0.90190	0.92723	0.93178
Recall	0.89040	0.92703	0.93176

TABLE III
MAIN RESULTS ON THE TESTING SET OF THE SUBTASK 1 WITH THE BERT-BASED MODELS.

Model	1	2	3	4
F1-macro	0.46458	0.49088	0.62265	0.62345
F1-weighted	0.47063	0.49646	0.62597	0.62676
Accuracy	0.57306	0.49646	0.66897	0.66966
Precision	0.54565	0.55434	0.60833	0.60875
Recall	0.99812	0.9990	0.99437	0.99500

TABLE IV
MAIN RESULTS ON THE TESTING SET OF THE SUBTASK 1 WITH THE GPT-BASED MODELS.

Model	5	6	7
F1-macro	0.83272	0.71571	0.71588
F1-weighted	0.83314	0.71759	0.71776
Accuracy	0.83442	0.73548	0.73571
Precision	0.84168	0.80296	0.80365
Recall	0.83442	0.73548	0.73548

performance, surpassing the F1-macro score of the previous best model and displaying a considerable advantage over the other GPT-2 models.

ACKNOWLEDGMENT

The authors wish to thank the support of the Instituto Politécnico Nacional (COFAA, SIP-IPN, Grant SIP 20230140) and the Mexican Government (CONAHCyT, SNI).

REFERENCES

[1] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving Language Understanding by Generative Pre-Training," 2018.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language Models are Unsupervised Multitask Learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language Models are Few-Shot Learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[4] OpenAI, "GPT-4 Technical Report," *ArXiv*, vol. abs/2303.08774, 2023, unpublished.

[5] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "PaLM: Scaling Language Modeling with Pathways," *arXiv preprint arXiv:2204.02311*, 2022, unpublished.

[6] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971*, 2023, unpublished.

[7] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé *et al.*, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," *arXiv preprint arXiv:2211.05100*, 2022, unpublished.

[8] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending Against Neural Fake News," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] A. Uchendu, T. Le, K. Shu, and D. Lee, "Authorship Attribution for Neural Text Generation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 8384–8395.

[10] M. Weiss, "Deepfake Bot Submissions to Federal Public Comment Websites Cannot Be Distinguished from Human Submissions," *Technology Science*, vol. 2019121801, 2019.

[11] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359*, 2021, unpublished.

[12] D. Yan, M. Fauss, J. Hao, and W. Cui, "Detection of AI-generated Essays in Writing Assessment," *Psychological Testing and Assessment Modeling*, vol. 65, no. 2, pp. 125–144, 2023.

[13] A. M. Sarvazyan, J. Á. González, M. Franco Salvador, F. Rangel, B. Chulvi, and P. Rosso, "Overview of AuTexTification at IberLEF 2023: Detection and Attribution of Machine-Generated Text in Multiple Domains," in *Procesamiento del Lenguaje Natural*, Jaén, Spain, sep 2023, in press.

[14] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, "How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection," *arXiv preprint arXiv:2301.07597*, 2023, unpublished.

[15] Y. Ma, J. Liu, F. Yi, Q. Cheng, Y. Huang, W. Lu, and X. Liu, "AI vs. Human–Differentiation Analysis of Scientific Content Generation," *arXiv preprint arXiv:1911.00650*, 2023, unpublished.

[16] F. Mireshghallah, J. Mattern, S. Gao, R. Shokri, and T. Berg-Kirkpatrick, "Smaller Language Models are Better Black-box Machine-Generated Text Detectors," *arXiv preprint arXiv:2305.09859*, 2023, unpublished.

[17] H. Zhan, X. He, Q. Xu, Y. Wu, and P. Stenetorp, "G3Detector: General GPT-Generated Text Detector," *arXiv preprint arXiv:2305.12680*, 2023, unpublished.

[18] S. Mitrović, D. Andreoletti, and O. Ayoub, "ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text," *arXiv preprint arXiv:2301.13852*, 2023, unpublished.

[19] Y. Li, Q. Li, L. Cui, W. Bi, L. Wang, L. Yang, S. Shi, and Y. Zhang, "Deepfake Text Detection in the Wild," *arXiv preprint arXiv:2305.13242*, 2023, unpublished.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 2019, pp. 4171–4186.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019, unpublished.

[22] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451, unpublished.

[23] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," *arXiv preprint arXiv:2006.03654*, 2020, unpublished.