GPT-2 versus GPT-3 and Bloom: LLMs for LLMs Generative Text Detection

Fernando Aguilar-Canto[1,†], Marco Cardoso-Moreno[1,*,†], Diana Jiménez[1,†] and Hiram Calvo[1]

[1]*Computational Cognitive Sciences Laboratory, Center for Computing Research, Instituto Politécnico Nacional, Mexico City 07700, Mexico*

Abstract

With the advent and proliferation of advanced Large Language Models (LLMs) such as BLOOM, GPT series, and ChatGPT, there is a growing concern regarding the potential misuse of this technology. Consequently, it has become imperative to develop machine learning techniques that can discern whether a given text has been generated by an LLM or authored by a human. In this paper, we present our approach in the AuTexTification shared task, where we fine-tuned BERT-based models and GPT-2 Small. Remarkably, GPT-2 Small achieved the highest F1-macro score in the validation set, prompting us to evaluate its performance on the testing set. We achieved an F1-macro score of 0.74134, securing the third position in the benchmark. Furthermore, we extended our fine-tuning efforts to the model attribution subtask, yielding a F1-macro score of 0.52282.

Keywords

Generative Text Detection, Large Language Models (LLMs), Model Attibution, AuTexTification

1. Introduction

In recent years, numerous companies and research centers have introduced a variety of Large Language Models (LLMs) to the field, including the Generative Pre-trained Transformer (GPT) [1], GPT-2 [2], ChatGPT, GPT-3 [3], Pathways Language Model (PaLM) [4], GPT-4 [5], BLOOM [6], and others. While the development of LLMs has achieved remarkable success, it has also raised ethical and public concerns regarding their potential misuse in generating and disseminating fake news or inaccurate information [7, 3, 8], spam [9], propaganda [10], and even facilitating academic cheating [11].

In light of these concerns, it becomes crucial to develop effective technologies for the detection and classification of text generated by LLMs, to mitigate the consequences associated with the

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

†These authors contributed equally.

pherjev@gmail.com (F. Aguilar-Canto); mcardosom2021@cic.ipn.mx (M. Cardoso-Moreno); dianaljl.99@gmail.com (D. Jiménez); hcalvo@cic.ipn.mx (H. Calvo)

https://github.com/Pherjev (F. Aguilar-Canto); https://www.cic.ipn.mx/index.php/francisco-hiram-calvo-castro (H. Calvo)

0000-0002-7352-3182 (F. Aguilar-Canto); 0009-0001-1072-2985 (M. Cardoso-Moreno); 0000-0002-3326-557X (D. Jiménez); 0000-0003-2836-2102 (H. Calvo)

aforementioned issues. However, the task of classifying generated text from LLMs, such as GPT-3, has proven to be no better than random when performed by trained human evaluators [12]. This highlights the need to leverage LLMs themselves to aid in the classification of generated text.

In this paper, we evaluate multiple LLMs performance in differentiating machine-generated text from human-written text, as well as in the area of model attribution for LLMs. To accomplish this, we utilize the benchmark proposed by Sarvazyan *et al.* [13], named AuTexTification (Automated Text Identification), which was presented in the context of the IberLEF 2023 [14]. The benchmark includes two subtasks: the first involves identifying whether a given text is human or machine-generated, while the second requires performing model attribution.

This paper is structured as follows: Section 2 provides an overview of the related works in the field. Section 3 introduces the methods and models implemented in our study. Section 4 presents the empirical results obtained from our experiments. In Section 5, we discuss and analyze the results in detail. Finally, Section 6 offers a conclusion summarizing our observations and suggestions for potential future research directions.

2. Related Works

High-quality text generation has emerged as a recent development, leading to a relatively limited number of classifiers specifically designed for this task in recent language models. However, it is important to note that text generation itself is not a novel task, as evidenced by earlier studies such as [15, 16]. Despite the historical context, recent Large Language Models (LLMs) present unique challenges in terms of classification, as highlighted by studies like [12]. These studies have demonstrated the difficulties in accurately classifying text generated by contemporary LLMs. In this literature review, we discuss the related works implemented on recent LLMs: GPT-2, GPT-3, ChatGPT, and other contemporary models. Most papers focus on identifying the machine-generated text, while model attribution has barely been studied [17].

2.1. GPT-2

Different Machine/Deep Learning methods have been proposed for identifying machine-generated texts [18]. When it comes to text generated by GPT-2, Solaiman *et al.* [19] utilized classical machine learning classifiers such as logistic regression. Additionally, Gehrmann *et al.* [20] introduced Giant Language model Test Room (GLTR), a set of baseline statistical methods for analyzing statistical differences between human and machine-generated text. Their observations revealed that humans tend to use rare words more frequently compared to GPT-2-generated text.

LLMs have also been employed to detect text generated by other language models. In the case of GPT-2, Solaiman *et al.* [19] utilized zero-shot classification with GPT-2 o GLOVER [7], but did not achieve better comparative results than classical approaches. Fine-tuned LLMs have demonstrated better performance in this task. For GPT-2, Ippolito *et al.* [21] proposed the usage of human evaluators and the BERT model [22]. In contrast, Zellers *et al.* [7] found that their model, GLOVER, was more appropriate for this task compared to BERT and detectors like fastText [23]. Other authors, such as Uchendu *et al.* [8], also utilized GLOVER.

Another commonly used LLM for detecting generated content is RoBERTa [24]. Solaiman *et al.* [19] achieved approximately 95% accuracy in detecting web pages generated by GPT-2 XL using RoBERTa. In this study, they also employed GPT-2 models to identify content generated by other GPT-2 models of varying sizes. The findings indicated that a larger GPT-2 model can identify content generated by a smaller model, but the reverse relationship does not hold. Overall, the study concluded that a RoBERTa model of the same capacity as GPT-2 is better suited for the task at hand. In the case of machine-generated tweets, Fagni *et al.* [25] found that RoBERTa outperformed classical classifiers, Convolutional Neural Networks (CNNs), Recurrent Networks, and other LLMs such as BERT, DistilBERT [26], and XLNet [27]. Similar results were achieved by Tourille *et al.* [28] and Kumarage *et al.* [29] in the same context. RoBERTa is also utilized in the task of detecting technically generated text [30].

For languages other than English, there are few works. For instance, Harrag *et al.* [31] utilized AraBERT [32] for classifying Arabic GPT-2 machine-generated texts versus human-written texts.

2.2. GPT-3

In the case of GPT-3, researchers have explored classical approaches, including feature-based identification methods [33]. One notable method proposed for identifying text generated by GPT-3 is DetectGPT [34], which utilizes a statistical criterion. In certain tasks, this criterion has shown superior performance compared to large models like RoBERTa.

Uchendu *et al.* [8] introduced TuringBench, a benchmark that encompasses various LLMs such as GROVER, GPT-2, GLTR, BERT, and RoBERTa, for the purpose of classifying human-written versus machine-generated text. Regarding GPT-3, the GPT-2 classifier achieved an F1-score of only 0.5293, while the best-performing model (BERT) attained a higher F1-score of 0.7944.

2.3. Beyond GPT-3: ChatGPT and GPT-4

Other recent LLMs, including GPT-3.5-turbo and ChatGPT, have also been subject to study. Various classifiers have been proposed, ranging from Random Forest with stylometric features [35], XGBoost with feature extraction [36], RoBERTa [37, 38, 39, 40], DistilBERT [41], to OPT-125M [39].

Combining different LLMs has also been explored. Wang *et al.* [42] employed RoBERTa, GLTR, and selected features to identify mixed content generated by ChatGPT, GPT-3, GLOOM, and others. Li *et al.* [43] used GLTP, fastText, LongFormer [44], and DetectGPT to detect text produced by LLMs such as GPT-3.5, OPT [45], and LLaMA [46].

However, despite the efforts made to detect content generated by modern LLMs like ChatGPT, Pegoraro *et al.* [47] conducted an analysis revealing that none of the modern proposals are capable of accurately identifying text generated by ChatGPT. In the case of benchmarks like MGTBench [48], the ChatGPT detector proposed by Guo *et al.* [37] demonstrated superior performance compared to other methods evaluated, although it can still be evaded with minor perturbations.

Few models have considered GPT-4 identification. Works by Zaitsu and Jin [35] and Zhang *et al.* [40] have included GPT-4 in their investigations.

3. Methodology

In our study, we compared the following LLMs for the classification of machine-generated versus human-generated texts:

1. BERT [22]
2. RoBERTa [24]
3. XLM-RoBERTa [49]
4. DeBERTA [50]
5. GPT-2 [2]

The selection of BERT-based models was based on their performance in similar tasks, as mentioned in the literature review of the previous section. As for GPT-2, we chose this model because it is one of the largest openly available models.

We employed the default hyperparameters provided by HuggingFace for each model. The evaluation metrics we reported include F1-macro, F1-micro, F1-weighted, and accuracy. In this task, the F1-macro metric is considered the most important [13]. All models were fine-tuned using the training set partition of the provided dataset and the HuggingFace framework. For the GPT-2 models, a dense layer was added to the model's output for classification purposes. Hyperparameter optimization was not implemented, and default hyperparameters provided by HuggingFace were utilized instead.

3.1. Dataset

The AuTexTification dataset [51] comprises a collection of machine-generated texts produced by the following models:

A : BLOOM-1b7
B : BLOOM-3b
C : BLOOM-7b1
D : GPT-3 Babbage
E : GPT-3 Curie
F : GPT-3 DaVinci-003

For the first subtask, the training set consists of 33,845 text samples labeled as either machine-generated or human. The testing set for the second subtask comprises 22,416 machine-generated text samples, each labeled with one of the letters A-F corresponding to the model that generated the text. The dataset exhibits low-class imbalance, with an entropy ratio ($H/\log k$) of approximately 0.99996 for subtask 1 and 0.99975 for subtask 2, where k represents the number of classes and H denotes the Shannon entropy. The texts in both subtasks are short, with a maximum of 98 words for subtask 1 and 97 words for subtask 2. The fixed testing set for subtask 1 consists of 21,832 samples, while the testing set for subtask 2 comprises 5,605 samples. It is worth noting that the AuTexTification dataset provides separate datasets for English and Spanish languages, and in this study, we focused solely on the English subtasks.

Table 1
Main results on validation set.

Model	Epochs	F1-macro	F1-micro	F1-weighted	Accuracy
BERT-base-cased	6	0.90462	0.90471	0.90451	0.90471
BERT-base-uncased	3	0.88145	0.88181	0.88120	0.88181
RoBERTa-base	3	0.91277	0.91284	0.91268	0.91284
XLM-RoBERTa-base	3	0.89376	0.89408	0.89355	0.89408
DeBERTa-base	3	0.90903	0.90914	0.90890	0.90914
GPT-2 Small	1	**0.91668**	**0.91668**	**0.91668**	**0.91668**

4. Results

In [13], the quantitative results reported for this submision (CIC-IPN-CsCog group) in the AuTexTification benchmark are based on the F1-macro score. Confidence intervals for these scores are also provided in the paper. However, other metrics, such as accuracy, F1-micro, and others, were not reported in the paper, and therefore, we included them in our analysis.

4.1. Subtask 1: Human or Generated

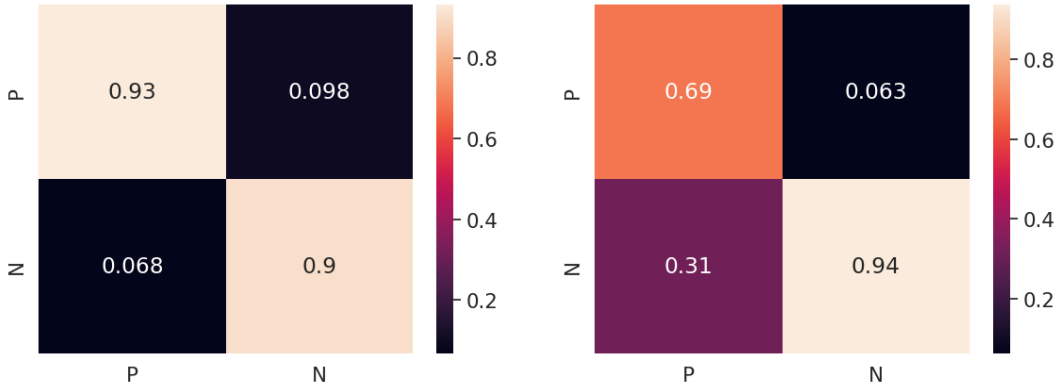
The first subtask of the AuTexTification dataset involves distinguishing between texts generated by LLMs and those written by humans. To evaluate the performance of the models, we randomly split 20% of the original training set as a validation set. We chose to train most models for 3 epochs, except for BERT-base-cased where we used 6 epochs to explore different results. Due to the computational cost associated with the GPT-2 model, we trained it for a single epoch. All models achieved similar metrics, with F1-scores and accuracy around 90%. However, GPT-2 outperformed the other models in various metrics, including F1-macro, F1-micro, F1-weighted, and accuracy. The detailed numerical results on the validation set are presented in Table 1. Furthermore, the confusion matrix of the best-performing model, GPT-2, is visualized in Figure 1a.

Subsequently, we proceeded to evaluate the performance of the GPT-2 model on the provided training set from the AuTexTification dataset [51]. The main results of both subtasks are summarized in Table 2. Additionally, Figure 1b presents the confusion matrix of the GPT-2 model specifically for subtask 1 in the testing set.

In subtask 1, the GPT-2 model achieved an F1-macro score of 0.74134, with a confidence interval of (73.53, 74.72). It is worth noting that our proposed model secured the third position in the original AuTexTification benchmark as described in [13]. Upon examining the results, we observe that the fine-tuned GPT-2 model exhibits a relatively high recall but a low precision in the testing set, leading to a considerable number of false negatives.

4.2. Subtask 2: Model attribution

In the second subtask of the AuTexTification benchmark, we utilized the GPT-2 model for classifying text generated by specific models, known as model attribution. It is important to note that for this subtask, we fine-tuned the GPT-2 model independently without transferring



(a) Validation test.

(b) Testing set.

Figure 1: Confusion matrices of the GPT-2 model applied on the validation (a) and testing set (b) of the subtask 1. The rows refer to the ground truth while the columns represents the predicted values. P stands for the positive class (human) whereas N stands for the negative class (generated).

Table 2
Main results on the testing set in both subtasks.

Subtask	F1-macro	F1-micro	F1-weighted	Accuracy
1	0.74134	0.74286	0.75559	0.74286
2	0.51988	0.51632	0.52110	0.51632

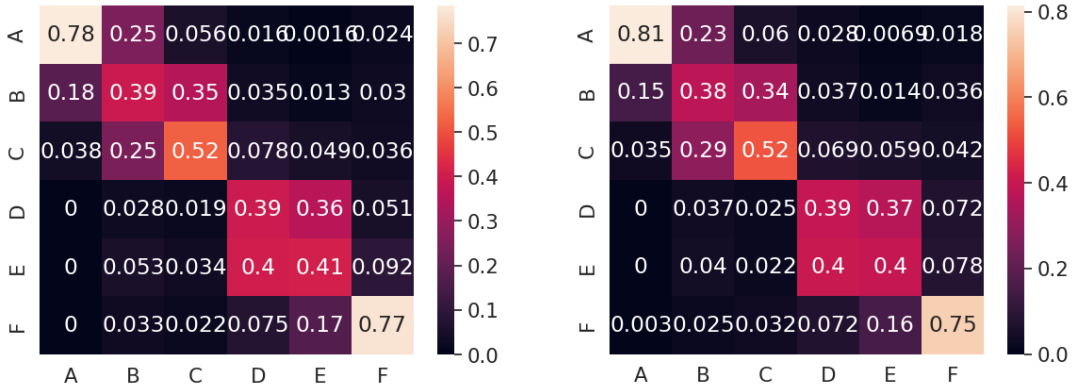
Table 3
Main results on the validation set of the subtask 2.

Model	Epochs	F1-macro	F1-micro	F1-weighted	Accuracy
GPT-2	1	0.52285	0.52230	0.52669	0.52230

any knowledge from the first subtask, according to the benchmark rules. The main results obtained on the validation set are provided in Table 3. Additionally, Figure 2a illustrates the confusion matrix generated by the GPT-2 model for the second subtask in the validation set.

In contrast to the first subtask, the results obtained in the testing set for the second subtask were quite similar. Table 2 presents the numerical results in the testing set, while Figure 2b depicts the confusion matrix for the same partition.

The GPT-2 model demonstrated the ability to accurately differentiate between models A (BLOOM-1b7), B-C (BLOOM-3b - BLOOM-7b1), D-E (GPT-3 Babbage - GPT-3 Curie), and F (GPT-3 DaVinci-003). However, most of the errors occurred in confusions between models B-C and D-E. Overall, the distinction between BLOOM and GPT-3 models was achieved correctly. It is worth noting that as reported in [13], this particular implementation of GPT-2 did not rank among the top models for this specific subtask, securing the 26th position. Nevertheless, it is possible to trace a hierarchical clustering of the models (figure 3).



(a) Validation test.

q

(b) Testing set.

Figure 2: Confusion matrices of the GPT-2 model applied on the validation (a) and testing set (b) of the subtask 2. The rows refer to the ground truth while the columns represents the predicted values. Letters A-F stands for the labels of the models.

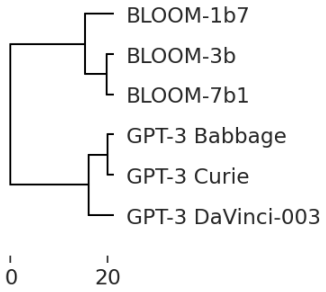


Figure 3: Hierarchical clustering of the models using the rule $D_{ij} = 2/(M_{ij} + M_{ji} + \varepsilon)$ where M is the confusion matrix.

5. Discussion

The results obtained in this article suggest that the GPT-2 model is suitable for identifying machine-generated content produced by GPT-3 and BLOOM models. However, it may not be the most effective approach for model attribution. Surprisingly, the smaller GPT-2 model was able to classify content generated by larger models, challenging the findings of the previous study of Solaiman *et al.* [19], where smaller GPT-2 models struggled to identify content generated by larger GPT-2 models.

Another point of contention is the effectiveness of BERT-based classifiers compared to GPT models, as discussed in [19]. In this study, GPT-2 outperformed BERT, RoBERTa, XLM-RoBERTa, and DeBERTa in the task of identifying machine-generated content. However, it should be noted that the comparison with RoBERTa was only made when the number of parameters in the GPT-2 model was similar, and a direct comparison between the two models was not performed.

It is important to mention that the AuTexTification dataset used in this study consists of

samples with less than 100 words, which is different from other approaches such as Yan *et al.* [11] where the classifier input is an essay. Therefore, this approach focuses more on short pieces of text rather than complete texts.

6. Conclusions

In this paper, we propose the use of fine-tuned GPT-2 models for both identifying machine-generated content by language models (LLMs) and the model attribution subtask. Given the widespread use of modern LLMs, both subtasks are crucial for monitoring and regulating the potential misuse of this technology.

To accomplish this, we compared the performance of BERT, RoBERTa, XLM-RoBERTa, DeBERTa, and GPT-2 models in the first subtask. Interestingly, the fine-tuned GPT-2 small model outperformed the other models, achieving a high F1-macro score of 0.91668 on the validation set. In the testing set, this model achieved a F1-macro score of 0.74134.

Furthermore, we applied the GPT-2 model to the second subtask, which involves model attribution. In this task, the GPT-2 model obtained a F1-macro score of 0.51988.

These results demonstrate the effectiveness of fine-tuned GPT-2 models for both identifying machine-generated content and performing model attribution. The utilization of GPT-2 small model showed promising performance, highlighting its potential for controlling and regulating the use of modern LLMs.

6.1. Further reseach

While this study focused on using the GPT-2 small model, there are other LLMs that can be considered for the same task, such as GPT-2 large, GPT-3, ChatGPT, and GPT-4. Future research can explore the performance of these models in identifying machine-generated content and model attribution. In addition, it must be explained the gap between the validation set F1-score and testing set F1-score in the subtask 1.

Furthermore, this study does not provide an explicit explanation for why GPT-2 small outperformed the BERT-based models. Exploring the explainability of these results would be an interesting avenue for further research. Understanding the factors that contribute to the internal decisions of the tested models can provide insights into how humans can be better prepared to distinguish between bot-generated and human-generated content.

Acknowledgments

The authors wish to thank the support of the Instituto Politécnico Nacional (COFAA, SIP-IPN, Grant SIP 20230140) and the Mexican Government (CONACYT, 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 1 (2019) 9.
[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 33 (2020) 1877–1901.
[4] 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).
[5] OpenAI, GPT-4 Technical Report, ArXiv abs/2303.08774 (2023).
[6] 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).
[7] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi, Defending Against Neural Fake News, Advances in Neural Information Processing Systems 32 (2019).
[8] A. Uchendu, T. Le, K. Shu, 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.
[9] M. Weiss, Deepfake Bot Submissions to Federal Public Comment Websites Cannot Be Distinguished from Human Submissions, Technology Science 2019121801 (2019).
[10] 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).
[11] D. Yan, M. Fauss, J. Hao, W. Cui, Detection of AI-generated Essays in Writing Assessment, Psychological Testing and Assessment Modeling 65 (2023) 125–144.
[12] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, N. A. Smith, All that's' human'is not gold: Evaluating human evaluation of generated text, arXiv preprint arXiv:2107.00061 (2021).
[13] A. M. Sarvazyan, J. Á. González, M. Franco Salvador, F. Rangel, B. Chulvi, 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, 2023.
[14] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, Procesamiento del Lenguaje Natural 71 (2023).
[15] W. C. Mann, C. M. Matthiessen, Nigel: A Systemic Grammar for Text Generation., Technical Report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST, 1983.
[16] K. McKeown, Text generation, Cambridge University Press, 1992.
[17] E. Merkhofe, D. Chaudhari, H. S. Anderson, K. Manville, L. Wong, J. Gante, Machine Learning Model Attribution Challenge, arXiv preprint arXiv:2302.06716 (2023).
[18] G. Jawahar, M. Abdul-Mageed, V. Laks Lakshmanan, Automatic Detection of Machine Generated Text: A Critical Survey, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 2296–2309.
[19] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger,

J. W. Kim, S. Kreps, et al., Release Strategies and the Social Impacts of Language Models, arXiv preprint arXiv:1908.09203 (2019).
[20] S. Gehrmann, S. Harvard, H. Strobelt, A. M. Rush, GLTR: Statistical Detection and Visualization of Generated Text, in: Proceedings of System Demonstrations, 2019, pp. 111–116.
[21] D. Ippolito, D. Duckworth, C. Callison-Burch, D. Eck, Automatic detection of generated text is easiest when humans are fooled, arXiv preprint arXiv:1911.00650 (2019).
[22] J. Devlin, M.-W. Chang, K. Lee, 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.
[23] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information, Transactions of the association for computational linguistics 5 (2017) 135–146.
[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv preprint arXiv:1907.11692 (2019).
[25] T. Fagni, F. Falchi, M. Gambini, A. Martella, M. Tesconi, TweepFake: About detecting deepfake tweets, Plos one 16 (2021) e0251415.
[26] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).
[27] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, XLNet: Generalized Autoregressive Pretraining for Language Understanding, Advances in Neural Information Processing Systems 32 (2019).
[28] J. Tourille, B. Sow, A. Popescu, Automatic Detection of Bot-generated Tweets, in: Proceedings of the 1st International Workshop on Multimedia AI against Disinformation, 2022, pp. 44–51.
[29] T. Kumarage, J. Garland, A. Bhattacharjee, K. Trapeznikov, S. Ruston, H. Liu, Stylometric Detection of AI-Generated Text in Twitter Timelines, arXiv preprint arXiv:2303.03697 (2023).
[30] J. Rodriguez, T. Hay, D. Gros, Z. Shamsi, R. Srinivasan, 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, 2022, pp. 1213–1233.
[31] F. Harrag, M. Dabbah, K. Darwish, A. Abdelali, BERT Transformer model for Detecting Arabic GPT2 AutoGenerated Tweets, in: Proceedings of the Fifth Arabic Natural Language Processing Workshop, 2020, pp. 207–214.
[32] W. Antoun, F. Baly, H. Hajj, AraBERT: Transformer-based Model for Arabic Language Understanding, in: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020, pp. 9–15.
[33] L. Fröhling, A. Zubiaga, Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover, PeerJ Computer Science 7 (2021) e443.
[34] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature, arXiv preprint arXiv:2301.11305 (2023).
[35] W. Zaitsu, M. Jin, Distinguishing ChatGPT (-3.5,-4)-generated and human-written papers

through Japanese stylometric analysis, arXiv preprint arXiv:2304.05534 (2023).
[36] R. Shijaku, E. Canhasi, Chatgpt generated text detection, ResearchGate (2023).
[37] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, Y. Wu, How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection, arXiv preprint arXiv:2301.07597 (2023).
[38] Y. Ma, J. Liu, F. Yi, Q. Cheng, Y. Huang, W. Lu, X. Liu, AI vs. Human–Differentiation Analysis of Scientific Content Generation, arXiv preprint arXiv:1911.00650 (2023).
[39] F. Mireshghallah, J. Mattern, S. Gao, R. Shokri, T. Berg-Kirkpatrick, Smaller Language Models are Better Black-box Machine-Generated Text Detectors, arXiv preprint arXiv:2305.09859 (2023).
[40] H. Zhan, X. He, Q. Xu, Y. Wu, P. Stenetorp, G3Detector: General GPT-Generated Text Detector, arXiv preprint arXiv:2305.12680 (2023).
[41] S. Mitrović, D. Andreoletti, 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).
[42] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, C. Whitehouse, O. M. Afzal, T. Mahmoud, A. F. Aji, et al., M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection, arXiv preprint arXiv:2305.14902 (2023).
[43] Y. Li, Q. Li, L. Cui, W. Bi, L. Wang, L. Yang, S. Shi, Y. Zhang, Deepfake Text Detection in the Wild, arXiv preprint arXiv:2305.13242 (2023).
[44] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The Long-Document Transformer, arXiv preprint arXiv:2004.05150 (2020).
[45] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al., OPT: Open Pre-trained Transformer Language Models, arXiv preprint arXiv:2205.01068 (2022).
[46] 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).
[47] A. Pegoraro, K. Kumari, H. Fereidooni, A.-R. Sadeghi, To ChatGPT, or not to ChatGPT: That is the question!, arXiv preprint arXiv:2304.01487 (2023).
[48] X. He, X. Shen, Z. Chen, M. Backes, Y. Zhang, MGTBench: Benchmarking Machine-Generated Text Detection, arXiv preprint arXiv:2303.14822 (2023).
[49] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, 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.
[50] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced BERT with Disentangled Attention, arXiv preprint arXiv:2006.03654 (2020).
[51] A. Sarvazyan, J. Ángel González, M. Franco, F. M. Rangel, M. A. Chulvi, P. Rosso, AuTexTification Dataset (Full data), 2023. URL: https://doi.org/10.5281/zenodo.7956207. doi:10.5281/zenodo.7956207.