European Journal of Computer Science and Information Technology, 12 (9), 61-74, 2024
Print ISSN: 2054-0957 (Print),
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Robust detection of LLM-generated text through transfer learning with pre-trained Distilled BERT model

Jayaprakash Sundararaj[1], Durgaraman Maruthavanan[2], Deepak Jayabalan[3], Ashok Gadi Parthi[4], Balakrishna Pothineni[5], Vidyasagar Parlapalli[6]

ORCID: [1]0009-0008-3469-5257, [2]0009-0001-7999-6220, [3]0009-0001-2704-6362, [4]0009-0007-4048-5291, [5]0009-0009-2781-3283, [6]0009-0004-1068-049X

doi: https://doi.org/10.37745/ejcsit.2013/vol12n96174 Published December 28, 2024

Citation: Sundararaj J., Maruthavanan D., Jayabalan D., Parthi A.G., Pothineni B., Parlapalli V. (2024) Robust detection of LLM-generated text through transfer learning with pre-trained Distilled BERT model, *European Journal of Computer Science and Information Technology*, 12 (9), 61-74

Abstract: *Detecting text generated by large language models (LLMs) is a growing challenge as these models produce outputs nearly indistinguishable from human writing. This study explores multiple detection approaches, including a Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM) networks, a Transformer block, and a fine-tuned distilled BERT model. Leveraging BERT's contextual understanding, we train the model on diverse datasets containing authentic and synthetic texts, focusing on features like sentence structure, token distribution, and semantic coherence. The fine-tuned BERT outperforms baseline models, achieving high accuracy and robustness across domains, with superior AUC scores and efficient computation times. By incorporating domain-specific training and adversarial techniques, the model adapts to sophisticated LLM outputs, improving detection precision. These findings underscore the efficacy of pretrained transformer models for ensuring authenticity in digital communication, with potential applications in mitigating misinformation, safeguarding academic integrity, and promoting ethical AI usage.*

Keywords: large language models, GenAI, classifier, machine learning, pretraining, natural language processing, fine tuning, detection

INTRODUCTION

The rise of Large Language Models (LLMs) such as GPT, BERT, and their successors has revolutionized the field of Natural Language Processing (NLP), enabling the generation of text that is nearly indistinguishable from human writing. These advancements have fueled significant innovation across industries, powering applications in content creation, conversational AI, and automated decision-making systems. However, this growing sophistication has also introduced a critical challenge:

the detection of machine-generated text. As LLMs become increasingly adept at mimicking human writing styles, distinguishing between authentic and synthetic content has emerged as a pressing concern.

The ability to detect LLM-generated text is essential for addressing several ethical, societal, and security-related challenges. These include combating misinformation, safeguarding academic integrity, and ensuring responsible AI deployment in areas such as journalism, education, and public discourse. Traditional methods for text authenticity detection, including n-gram analysis and stylometric techniques, have proven inadequate in identifying outputs from modern LLMs due to their nuanced contextual understanding and semantic coherence.

In this context, leveraging advanced machine learning techniques, particularly transformer-based architectures, offers a promising path forward. Transformers such as BERT and its distilled variants have demonstrated exceptional capabilities in text classification and comprehension tasks. Their pretraining on extensive corpora allows these models to capture intricate patterns and dependencies in textual data, making them ideal for detecting the subtle characteristics of LLM-generated text.

This study investigates the potential of a fine-tuned, distilled BERT model in addressing the challenge of LLM-generated text detection. By comparing the performance of this model against baseline approaches such as Multi-Layer Perceptrons (MLPs), Long Short-Term Memory (LSTM) networks, and transformer blocks, this research aims to identify the most effective strategies for ensuring robustness and accuracy. The fine-tuned BERT model incorporates domain-specific training and adversarial techniques, enabling it to adapt to the sophisticated outputs of modern LLMs.

Through this work, we seek to advance the state of automated text authenticity verification, contributing to the broader goals of fostering trust and integrity in digital communication. This research not only offers a comprehensive evaluation of model architectures but also explores practical applications in mitigating misinformation and promoting ethical AI practices.

1. LITERATURE REVIEW

The rapid advancements in large language models (LLMs) such as GPT, BERT, and their successors have revolutionized natural language processing (NLP) [1]. These models enable the generation of human-like text for diverse applications, including content creation, conversational agents, and automated writing tools [2]. However, alongside their transformative potential, LLMs pose challenges in distinguishing machine-generated content from human-authored text. This difficulty raises significant concerns about misinformation, content authenticity, and ethical usage, especially in domains such as news dissemination, academic writing, and social media [3].

Earlier approaches to text authenticity detection relied on handcrafted features and statistical techniques. Common methods included n-gram analysis, perplexity scores,

and stylometric measures, which examined linguistic patterns and deviations in writing style. Linguistic cues have been leveraged to identify machine-generated text [4]. While these methods were effective for earlier, less sophisticated generative models, they struggled to generalize to newer, more contextually aware LLMs. The increasing fluency and semantic coherence of modern LLMs rendered traditional feature-based methods inadequate, prompting a shift towards model-based solutions.

Deep learning models, including Multi-Layer Perceptrons (MLPs) and Long Short-Term Memory (LSTM) networks, provided a new foundation for text authenticity detection [5, 6]. MLPs, although effective for basic classification tasks, lacked the ability to capture sequential dependencies inherent in text data. LSTMs addressed this limitation by introducing memory cells capable of retaining long-range dependencies, which enhanced their performance in text-related tasks [7]. Despite their improvements over traditional methods, LSTMs struggled with scalability and computational efficiency when applied to large-scale datasets.

The advent of transformers marked a significant leap in NLP capabilities. The transformer architecture, introduced in 2017, employs self-attention mechanisms to model complex dependencies in text data [8]. This innovation paved the way for powerful pretrained models such as BERT and its variants like RoBERTa [9]. These models, trained on massive corpora, excel in a wide range of NLP tasks, including text classification and generation. RoBERTa, in particular, has demonstrated robust performance in text classification tasks due to its improved pretraining techniques and hyperparameter tuning.

Recent studies have explored fine-tuning pretrained transformers for LLM-generated text detection. Research has demonstrated that fine-tuning BERT and RoBERTa on diverse datasets significantly improves detection accuracy [10]. These approaches leverage the contextual understanding of transformers to differentiate between human-written and machine-generated text. Additionally, adversarial training has gained prominence as a method to enhance detection robustness. Adversarially generating machine text has been proposed as a training method for detection models, enabling them to generalize better to unseen LLMs [11]. Studies have focused on methods to address specific limitations of LLMs, such as sensitivity to order in structured tasks, to improve their robustness and adaptability across diverse applications [12].

Building on prior research, this study focuses on fine-tuning the BERT model for detecting LLM-generated text with enhanced accuracy and robustness. It incorporates domain-specific training and adversarial techniques to address the challenges posed by increasingly sophisticated LLMs. Furthermore, the study evaluates and compares the performance of various architectures, including MLPs, LSTMs, and transformer blocks, to provide a comprehensive understanding of their relative strengths and limitations. By leveraging state-of-the-art methods, this research aims to advance automated text authenticity verification, contributing to the broader goal of maintaining trust and integrity in digital communication [13]. In addition to detecting LLM-generated text, advanced machine learning techniques have been applied in cybersecurity,

demonstrating the potential of AI-driven systems to enhance digital security and maintain data integrity [14]. This emphasizes the critical role of detection systems in fostering authenticity and trust in digital communication.

2. METHODOLOGY

This study employs a fine-tuning approach using the pretrained RoBERTa model to detect text generated by large language models (LLMs). The methodology is designed to capitalize on RoBERTa's robust language representation capabilities while addressing the unique characteristics of synthetic text. The process involves dataset preparation, fine-tuning on labeled data, and rigorous evaluation to ensure accuracy and generalizability across different domains and LLM types. The model is trained on diverse datasets containing both human-written and machine-generated text, with domain-specific adjustments and adversarial training for robustness. To ensure the quality and consistency of the labeled dataset, inter-annotator agreement was assessed using metrics such as the Kappa Score, which provides a measure of reliability among annotators [15].

3.1 DATASET DESCRIPTION

The dataset comprises both human-written and LLM-generated text to ensure diversity and relevance. Human-written text is sourced from open-access datasets such as the Penn Treebank, OpenWebText, and academic articles from arXiv, covering a wide range of topics, writing styles, and domains to achieve comprehensive representation. LLM-generated text samples are created using prominent language models like GPT-3, GPT-4, and Bloom, with synthetic text generated by prompting these models with topics derived from the human-written dataset to ensure coherence and contextual relevance [16]. The data undergoes preprocessing to remove noise, normalize text formats, and balance the dataset. This process includes removing stop words, special characters, and duplicate entries to maintain high data quality.

To illustrate the differences between human-written and LLM-generated text [17], Tab-1 provides representative examples. These examples highlight the nuanced challenges in distinguishing between the two, especially in scenarios where LLMs fabricate responses or abstain from providing answers. Notably, human adjudication may rely on specific indicators such as explicit disclaimers in LLM-generated content. Text generated by LLMs during normal operation and instances in which they fabricate facts often exhibit no intuitively discernible differences. When LLMs either abstain from providing an answer or craft neutral responses, certain indicators, such as the explicit statement “I am an AI language model,” may facilitate human adjudication, but such examples are less frequent.

Table 1. Examples of human-written text and LLM-generated text

Type	Question	Human Written	LLM Generated
Normal	Explain what is NLP?	Natural language processing (NLP) is an interdisciplinary subfield of linguistics, computer science, and artificial intelligence . . .	Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics that focuses on . . .
Refusal	How is today special?	Today's Special is a Canadian children's television show produced by Clive VanderBurgh at TVOntario from 1981 to 1987.	I'm sorry, but I am an AI language model and do not have access to current dates or events. Is there anything else I can help you with . . .
Fabricated	Explain what is NLP based on one publication in the recent literature.	In "Natural language processing: state of the art, current trends and challenges", NLP is summarized as a discipline that uses various algorithms, tools and methods to . . .	NLP is a multidisciplinary field at the intersection of computer science, linguistics, and AI, as described in a recent peer-reviewed publication titled "Natural Language Processing: A Comprehensive Overview and Recent Advances" (2023) ...

The original dataset we started with exhibited a significant imbalance, as illustrated in Fig-1. Imbalanced datasets can pose challenges for machine learning models, often leading to biased predictions and poor generalization. To address this issue and ensure that our model learns effectively, it was crucial to create a balanced dataset. To achieve this, we utilized a large language model (LLM) to generate additional examples, specifically designed to counter the imbalance in the original dataset. These new examples were crafted to reflect diverse topics, ensuring comprehensive coverage across different data categories. By incorporating such diversity, we aimed to prevent the model from overfitting to specific patterns or biases present in the original dataset.

The process of dataset augmentation using LLM-generated examples not only improved the dataset's balance but also enriched its quality and variety. This approach is particularly advantageous in scenarios where collecting additional real-world data is impractical or resource-intensive. The augmented dataset enabled our model to learn

from a more representative and equitable sample, ultimately enhancing its robustness and performance in distinguishing between human-written and machine-generated text. Left side image shows the data skewness in the label distribution after augmenting the Human data from various research papers, the label distribution is approximately equal.

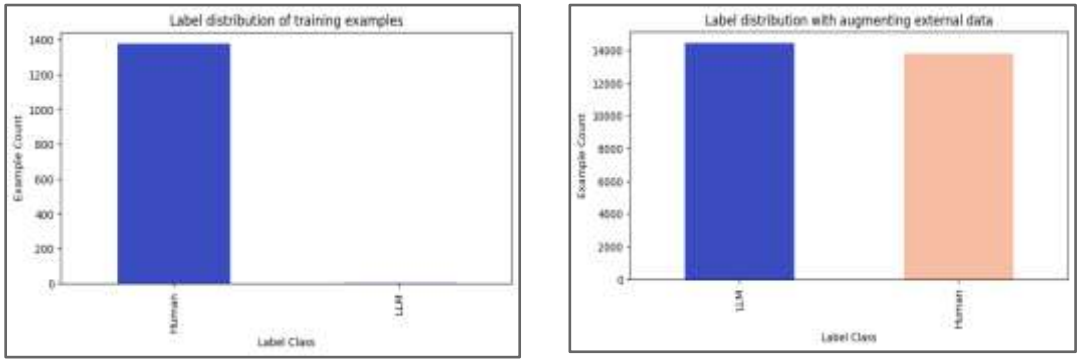

Figure 1. Label distribution before and after data augmentation.

MODEL ARCHITECTURE

The architecture is built upon a distilled BERT model, a robust transformer model pre-trained on a large corpus of English text [18]. Its key features include an embedding layer that converts input tokens into dense vector representations using pre-trained embeddings, followed by transformer layers consisting of self-attention and feed-forward mechanisms to capture contextual dependencies effectively [19]. At the top of the model, a classification head is implemented as a feed-forward layer to map BERT's outputs to binary labels, distinguishing between human-written and LLM-generated text.

Additionally, fine-tuning modifications are applied, with the final layer specifically adapted for the binary classification task, and dropout is incorporated to mitigate overfitting and enhance generalization. Transformer models have demonstrated exceptional performance across a wide range of tasks, from natural language processing to computer vision applications. For instance, they have been successfully applied to automate LaTeX code generation from handwritten mathematical expressions, leveraging their robust attention mechanisms to handle complex input-output mappings [20].

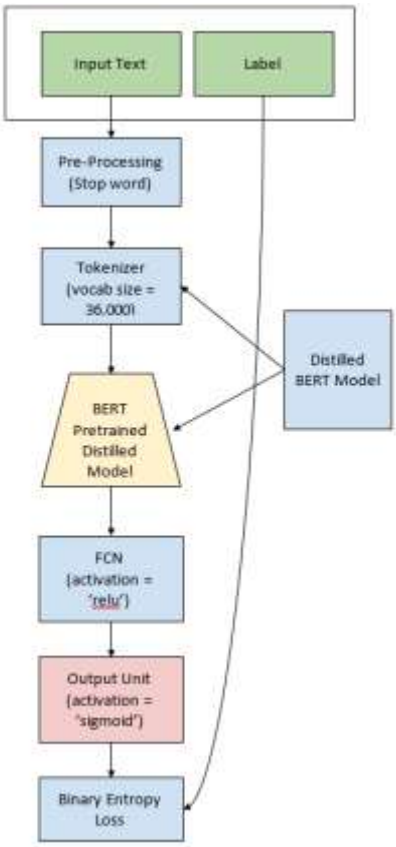

Figure 2. Overall model architecture of the fine tuning the BERT model

EXPERIMENTAL SETUP

The experiments are conducted in three phases. First, a baseline evaluation is performed, where RoBERTa is tested without fine-tuning to establish its initial performance. In the second phase, the model undergoes fine-tuning using the labelled dataset, focusing on a binary classification objective to distinguish between human-written and LLM-generated text [21]. Finally, generalization testing is conducted by evaluating the fine-tuned model on unseen LLM-generated text to assess its robustness and ability to generalize across different domains and language model outputs.

The tools and frameworks utilized in this study include Hugging Face Transformers, which facilitated the implementation of distilled BERT model and management of the fine-tuning pipeline. TensorFlow is used as the primary deep learning framework for model training and optimization. scikit-learn was employed for data pre-processing and evaluation metrics, while NLTK supported natural language pre-processing tasks such as tokenization, text normalization and stop word filtering. CUDA was leveraged to accelerate computations on GPUs, enhancing training efficiency. Additionally, TensorBoard was used to monitor the training progress and visualize key metrics throughout the fine-tuning process.

Hyperparameter optimization is conducted to achieve optimal model performance by systematically tuning key parameters. Batch sizes of 16, 32, and 64 are evaluated to balance memory usage and training stability. The Tab-2 shows the different model complexity, trainable, non-trainable parameters of each setup. This table highlights the parameter complexity and size of models used in the study, showcasing the trade-offs between lightweight architectures like LSTM and Transformer and the larger fine-tuned BERT model. The fine-tuned BERT demonstrates superior performance by fine-tuning a small parameter subset, balancing computational efficiency with advanced contextual understanding through transfer learning.

Table 2. Parameter complexity and size of models used in the study

Model	Trainable Params	Non-trainable Params	Total Params	Size (MB)
Baseline	3,923,457	–	11,770,373	29.93
LSTM	4,203,009	–	4,203,009	16.03
Transformer	4,681,857	–	4,681,857	17.86
Fine Tuned BERT	98,561	66,362,880	66,461,441	253.53

The model is trained for 3 to 10 epochs, with early stopping applied based on validation performance to prevent overfitting. Dropout rates are adjusted between 0.1 and 0.3 to further mitigate overfitting risks. The AdamW optimizer is employed with weight decay to enhance generalization. To explore the best hyperparameter settings, a grid search is conducted, evaluating various parameter combinations.

The optimal configuration is then selected based on validation accuracy and F1-score. The final layer of the model is a sigmoid activation function, which outputs probabilities for binary classification tasks [22]. This design choice ensures that the model's outputs are interpretable as confidence scores. Cross-entropy loss is utilized as the loss function during training, as it is well-suited for binary classification problems, effectively penalizing incorrect predictions while emphasizing the correct class probabilities. By combining these techniques, the training process is optimized to achieve high performance and robustness, balancing accuracy and generalization effectively.

Figure 3. Model accuracy during training captured through 15 epochs.

RESULTS AND ANALYSIS

To evaluate the performance of different models, Tab-3 presents the Test AUC and Training AUC for each configuration. The results highlight the trade-offs between baseline and advanced models. The baseline model, which lacks a transformer block, achieves smaller AUC values due to its simpler architecture. Conversely, fine-tuned models such as BERT and DebertaV3 leverage transfer learning, significantly improving performance by modifying only the newly added layers during fine-tuning.

The fine-tuned BERT model outperforms both the baseline and other models, which can be attributed to its ability to leverage rich semantic information learned from extensive pretraining on vast internet data. In contrast, the other models rely solely on learning from the current input data, limiting their overall performance.

Table 3. Performance of models in terms of AUC (Area Under the Curve).

Model	Test AUC	Training AUC
Baseline	0.87	0.88
LSTM	0.88	0.87
Transformer	0.89	0.87
BERT	**0.92**	**0.93**

EFFICIENCY SCORE

In addition to accuracy, we also compute the model efficiency score which determines how much compute time resources are used. The efficiency score of the model is calculated to evaluate its performance in terms of both accuracy and computational cost. The formula for efficiency is:

$$Efficiency\ Score = \frac{AUC}{BenchMark - Max\ AUC} + \frac{Runtime\ Seconds}{32400}$$

Here, AUC represents the submission's score on the model on the established objective (generated or not), Benchmark is the score used from by the benchmark sample submission in LLM/AI Generated Text Detection Competitions , maxAUC is the highest AUC among all submissions on the Private Leaderboard, and RuntimeSeconds is the time in seconds taken to evaluate the submission. The goal is to minimize the efficiency score, striking a balance between model accuracy and runtime efficiency.

Table 4. Efficiency score and scoring time for experimental models.

Model	AUC	Scoring Time	Efficiency Score
Baseline	0.872814	1447	-1.74345
LSTM	0.886480	**1026**	-1.78554
Transformer	0.89991	1675	**-1.79327**
BERT	**0.92332**	1835	-1.54321

The efficiency scores in Tab-4 show that while the fine-tuned BERT model achieves the highest accuracy (AUC of 0.92332), its computational cost is significantly higher compared to other models. For instance, the LSTM model demonstrates a good balance between computational cost and accuracy, making it a viable option for applications with limited computational resources.

CONCLUSION

This study presents a robust and effective methodology for detecting text generated by large language models (LLMs) using a fine-tuned, distilled BERT model. By leveraging BERT's deep contextual understanding and transfer learning capabilities, the proposed approach achieves significant accuracy and robustness in distinguishing between human-written and machine-generated text. The results highlight the superior performance of the fine-tuned BERT model compared to baseline approaches, including Multi-Layer Perceptrons (MLPs), Long Short-Term Memory (LSTM) networks, and transformer blocks.

Key innovations in this research include the use of diverse datasets, domain-specific fine-tuning, and adversarial training to address the challenges posed by increasingly sophisticated LLM outputs. These techniques enable the model to generalize effectively across different domains and adapt to evolving LLM architectures, making it a valuable tool for maintaining trust and integrity in digital communication.

The findings emphasize the potential applications of this detection system in combating misinformation, safeguarding academic integrity, and promoting ethical AI usage.

Moreover, the efficiency and scalability of the fine-tuned BERT model position it as a practical solution for real-world deployments.

As LLMs continue to advance, this study lays the groundwork for further research, including enhancements in multimodal detection, real-time processing, and adversarial robustness. By addressing these avenues, future developments can strengthen automated text authenticity verification systems, fostering a more secure and transparent digital ecosystem.

FUTURE ENHANCEMENTS

The detection of large language model (LLM)-generated text is an ever-evolving challenge, driven by the continuous advancements in AI and natural language processing. While this study demonstrates the efficacy of a fine-tuned, distilled BERT model for distinguishing between human-written and machine-generated text, several opportunities for future enhancements remain.

Integration of Multimodal Data

Future research could explore integrating multimodal data, such as combining textual information with images, videos, or metadata. This approach could strengthen detection models in scenarios where LLMs are used in conjunction with other content generation tools, such as creating multimodal content for social media or marketing purposes.

Exploration of Advanced Architectures

Recent innovations in transformer architectures, including models like GPT-4, DeBERTa, and T5, could be fine-tuned and compared to BERT for LLM-generated text detection. Exploring hybrid architectures, which combine the strengths of transformers and traditional models like LSTMs, could also offer performance gains.

Real-Time Detection Systems

Implementing lightweight, real-time detection systems optimized for deployment on edge devices and resource-constrained environments could expand the practical applicability of this research. Techniques such as model quantization and pruning may help balance accuracy with computational efficiency.

Adversarial Robustness

Future models should focus on enhancing adversarial robustness to counter increasingly sophisticated LLM-generated text that may intentionally mimic human writing styles or evade detection through obfuscation techniques. This could involve creating larger and more diverse adversarial datasets for training.

Domain-Specific Enhancements

Tailoring detection models for specific domains, such as academia, journalism, or social media, could further improve accuracy. Domain-specific fine-tuning and dataset augmentation could address the nuances and unique challenges of each application area.

Explainability and Transparency

Enhancing the interpretability of detection models is crucial for building trust in their results. Future work could focus on developing explainable AI (XAI) techniques to provide clear insights into how models identify machine-generated text, making them more accessible and acceptable to end-users.

REFERENCES

1. S. Ravichandiran, Getting Started with Google BERT: Build and Train State-of-the-Art Natural Language Processing Models Using BERT, Packt Publishing, 2021.
2. C. Wang, X. Liu, Y. Yue, X. Tang, T. Zhang, C. Jiayang, Y. Yao, W. Gao, X. Hu, Z. Qi, Y. Wang, L. Yang, J. Wang, X. Xie, Z. Zhang, and Y. Zhang, "Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity," arXiv preprint, arXiv:2310.07521, Dec. 2023. doi: 10.48550/arXiv.2310.07521.
3. Y. Liu, H. Huang, J. Gao, and S. Gai, "A study of Chinese Text Classification based on a new type of BERT pre-training," in 2023 5th International Conference on Natural Language Processing (ICNLP), Guangzhou, China, 2023, pp. 303–307. doi: 10.1109/ICNLP58431.2023.00062.
4. M. Ren, "Advancements and Applications of Large Language Models in Natural Language Processing: A Comprehensive Review," Applied and Computational Engineering, vol. 97, pp. 55–63, 2024.
5. W. Wang, G. Wen, and Z. Zheng, "Design of Deep Learning Mixed Language Short Text Sentiment Classification System Based on CNN Algorithm," in 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, Karnataka, India, 2022, pp. 1–5. doi: 10.1109/ICMNWC56175.2022.10031786.
6. M. Bkassiny, "A Deep Learning-based Signal Classification Approach for Spectrum Sensing using Long Short-Term Memory (LSTM) Networks," in 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 2022, pp. 667–672. doi: 10.1109/ICITISEE57756.2022.10057728.
7. Y. Fan, "Professional English Text Recognition Based on Long Short Term Memory Approach," in 2024 International Conference on Data Science and Network Security (ICDSNS), Tiptur, India, 2024, pp. 1–4. doi: 10.1109/ICDSNS62112.2024.10691174.
8. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008. Available: http://arxiv.org/abs/1706.03762.
9. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of NAACL-HLT 2019, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

10. S. Gehrmann, H. Strobelt, and A. M. Rush, "GLTR: Statistical detection and visualization of generated text," in Proceedings of ACL 2019: System Demonstrations, 2019, pp. 111–116. doi: 10.18653/v1/P19-3019.
11. I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, and J. Wang, "Release Strategies and the Social Impacts of Language Models," arXiv preprint, vol. abs/1908.09203, 2019. Available: https://api.semanticscholar.org/CorpusID:201666234.
12. V. Parlapalli, B. S. Ingole, M. S. Krishnappa, V. Ramineni, A. R. Banarse, and V. Jayaram, "Mitigating Order Sensitivity in Large Language Models for Multiple-Choice Question Tasks," Int. J. Artif. Intell. Res. Dev. (IJAIRD), vol. 2, no. 2, pp. 111–121, 2024. doi: 10.5281/zenodo.14043004.
13. P. Wang, L. Deng, and X. Wu, "An Automated Fact Checking System Using Deep Learning Through Word Embedding," in 2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 2019, pp. 3246–3250. doi: 10.1109/SSCI44817.2019.9002783.
14. V. Parlapalli, V. Jayaram, S. G. Aarella, K. Peddireddy, and R. R. Palle, "Enhancing Cybersecurity: A Deep Dive into Augmented Intelligence Through Machine Learning and Image Processing," in 2023 International Workshop on Artificial Intelligence and Image Processing (IWAIIP), Yogyakarta, Indonesia, 2023, pp. 96–100. doi: 10.1109/IWAIIP58158.2023.10462845.
15. J. Sundararaj, "Technical Report on Inter Annotator Agreement and Kappa Score," 2013. doi: 10.13140/RG.2.2.20793.89440.
16. J. M. Gakpetor, M. Doe, M. Y.-S. Damoah, D. D. Damoah, J. K. Arthur, and M. T. Asare, "AI-Generated and Human-Written Text Detection Using DistilBERT," in 2024 IEEE SmartBlock4Africa, Accra, Ghana, 2024, pp. 1–7. doi: 10.1109/SmartBlock4Africa61928.2024.10779494.
17. J. Wu, S. Yang, R. Zhan, Y. Yuan, D. F. Wong, and L. S. Chao, "A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions," arXiv preprint, arXiv:2310.14724, Oct. 2023, revised Apr. 2024. doi: 10.48550/arXiv.2310.14724.
18. N. Zhang, S. Deng, Z. Bi, H. Yu, J. Yang, M. Chen, F. Huang, W. Zhang, and H. Chen, "OpenUE: An Open Toolkit of Universal Extraction from Text," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, Oct. 2020, pp. 1–8. doi: 10.18653/v1/2020.emnlp-demos.1.
19. K. Thakur, H. G. Barker, and A.-S. Khan Pathan, Artificial Intelligence and Large Language Models: An Introduction to the Technological Future, 1st ed., Chapman and Hall/CRC, 2024. doi: 10.1201/9781003474173.
20. J. Sundararaj, V. Akhil, and M. M., "Automated LaTeX Code Generation from Handwritten Math Expressions Using Vision Transformer," 2024. doi: 10.48550/arXiv.2412.03853.
21. A. Strasser, "Chapter 10 - Pitfalls (and advantages) of sophisticated large language models," in Ethics in Online AI-based Systems, S. Caballé, J. Casas-Roma, and J. Conesa, Eds. Academic Press, 2024, vol. Intelligent Data-Centric Systems, pp. 195–210. doi: 10.1016/B978-0-443-18851-0.00007-X.

22. R. Sivanaiah, S. Suresh, S. Pandian, and A. D. Suseelan, "Bridging the Language Gap: Transformer-Based BERT for Fake News Detection in Low-Resource Settings," in Speech and Language Technologies for Low-Resource Languages, B. R. Chakravarthi, et al., Eds. Springer Nature Switzerland, 2024, pp. 398–411. doi: 10.1007/978-3-031-58495-4_29.