Small Language Models for Fake News Detection: A Green AI Approach

Anonymous Author(s)

Abstract

Artificial intelligence (AI) has become a valuable tool across a wide range of applications, including the detection of fake news on social media. However, as AI systems grow more complex, their energy demands during both training and deployment have increased substantially. In response, Green AI has emerged, promoting the development of models that are not only effective but also energy-efficient and environmentally sustainable. This study applies computational methods to pursue two primary goals. First, we aim to improve fake news detection using small language models, such as Bidirectional Encoder Representations from Transformers (BERT), a pretrained transformer model capable of capturing nuanced language and contextual meaning. Second, we evaluate lightweight variants of BERT that employ techniques like knowledge distillation and parameter sharing to reduce computational overhead without compromising performance. In experiments using two public datasets, we achieved accuracy rates of 99.47% and 99.83%, reducing the likelihood of misclassification by traditional machine learning methods by 86% and 96%, respectively. Moreover, we found that distilled versions of BERT can maintain high accuracy while requiring only half the runtime compared to the original BERT model when applied to small training datasets.

Keywords

Fake News, Detection, Artificial Intelligence, Transformer Models

1 Small Language Models for Fake News Detection: A Green AI Approach

AI is reshaping how news and information circulate online, from personalized feeds and automatic translations to algorithmically generated headlines. One area drawing particular attention is the detection of fake news on social media. Traditional fact-checking often comes too late, responding after false claims spread widely [31]. By contrast, AI-powered systems can analyze content in real time and flag misleading information before it circulates further [13]. Such systems preserve information integrity but also raise questions about the trade-offs between technological advancement and environmental sustainability.

A growing concern is the environmental cost of large-scale AI models. Training deep learning systems consumes vast amounts of electricity and produces substantial carbon emissions [22, 25]. Recent estimates suggest training models such as GPT-3 or GPT-4 requires thousands of megawatt-hours (MWh), with emissions comparable to burning hundreds or even thousands of tons of coal [10, 28]. Beyond training, everyday use also carries a heavy footprint: a single ChatGPT query uses about ten times the energy of a Google search, and industry-wide consumption may reach 300 gigawatt-hours (GWh) per day by 2027 [6, 12].

These challenges highlight the importance of *Green AI*, which emphasizes reducing energy consumption and resource use while maintaining strong performance [24]. In contrast, *Red AI* prioritizes accuracy gains through massive compute resources, often at high ecological cost. Applying Green AI principles to misinformation detection offers a path toward effective yet sustainable AI practices.

This study pursues two objectives. First, we evaluate small language models such as BERT [7], which, although modest in scale compared to trillion-parameter large language models (LLMs) like GPT-4 [20], remain highly effective for tasks requiring nuanced language understanding. Second, we investigate lightweight BERT variants, including DistilBERT and ALBERT, which reduce computational costs through distillation and parameter sharing while preserving accuracy [17, 23].

Our approach relies on linguistic analysis rather than external fact-checking (for a survey of detection methods, see [33]). BERT-based models classify news articles by recognizing patterns in tone, lexical choice, and syntax that often signal deceptive content [15]. In doing so, they demonstrate how transformer architectures can simultaneously advance misinformation detection and contribute to the broader sustainability goals of Green AI.

2 Literature Review

2.1 Machine Learning Approaches

2.1.1 Classical Machine Learning Models. Early efforts in fake news detection relied on classical machine learning models such as Support Vector Machines (SVM), Decision Trees, and Naïve Bayes classifiers. These models leverage basic linguistic features, including word frequency, sentence length, syntactic structures, and part-of-speech tags to distinguish between factual and fake news [1]. For example, fake news may include more sensational or emotionally charged words like “shocking” or “exposed,” which increase word frequency and emotional intensity. It also tends to use shorter, punchier sentences to create urgency and attract clicks. In contrast, factual news typically uses more neutral language, longer sentence structures, and formal syntactic patterns that reflect journalistic standards. Differences in part-of-speech usage, such as a higher frequency of adjectives and adverbs in fake news, have also been observed in prior studies (for a systematic review, see [5]).

2.1.2 Hybrid Machine Learning Models. While classical machine learning models are relatively easy to implement and have shown success in certain cases [15], they often struggle to capture the complexity of human language and the subtle distinctions between factual information and misinformation [11]. To enhance performance, researchers have introduced hybrid models that combine multiple components within a single architecture to leverage the strengths of each [14]. However, as misinformation becomes more context-dependent and linguistically nuanced, even hybrid approaches continue to face challenges in detecting the deeper semantic cues that often characterize fake news.

2.1.3 Word Embedding-Enhanced Models. To address some of these limitations, word embedding techniques such as Word2Vec and GloVe have been used to enrich text representations [21]. These embeddings convert words into vectors that capture semantic relationships based on co-occurrence patterns. For example, such models can recognize that "king" is to "queen" as "man" is to "woman," or that "apple" is more similar to "banana" than to "car." In the context of misinformation detection, a model may learn that terms like "breaking," "shocking," and "truth" frequently co-occur in fake news articles. Although embeddings have improved performance to some extent, they often fail to capture the contextual dependencies of words, limiting their effectiveness in complex classification tasks like fake news detection [30]. These limitations have prompted the adoption of more advanced models that can better account for word meaning in context.

2.2 BERT Models

Recent advancements in NLP, particularly the development of transformer-based models such as BERT, have significantly improved performance in text classification tasks, including fake news detection. BERT is built on a stacked self-attention mechanism, which enables the model to capture relationships between words across an entire sequence [29]. When fine-tuned on labeled fake news datasets, BERT learns to recognize subtle linguistic cues, such as emotional tone, exaggerated language, and structural irregularities, that often characterize misinformation [8].

Another feature that sets BERT apart from earlier models is its bidirectional architecture, which allows it to consider both the words that come before and after a given word. This enables a more comprehensive understanding of meaning, as words are interpreted within their full contextual environment [7]. In contrast, machine learning models often process text in a single direction, which can result in missed linguistic cues, especially in longer or more syntactically complex sentences [11]. By capturing surrounding context more effectively, BERT is better suited to detect the subtle and nuanced language features often present in misinformation [30].

Given these capabilities, this study evaluates the effectiveness of transformer-based models in detecting fake news, with a focus on BERT and its variants.

- **RQ1.** How effective are BERT-based models in detecting fake news compared to traditional machine learning models?

Although BERT is relatively lightweight compared to today's LLMs, further optimization is possible through the use of more efficient transformer architectures. Several alternatives have been introduced to reduce model size and computational demands without significantly compromising performance. For instance, DistilBERT compresses the original BERT architecture by approximately 40% using knowledge distillation, yet retains much of its accuracy [23]. ALBERT enhances efficiency through parameter sharing and factorized embeddings, significantly lowering memory usage while maintaining competitive performance [17]. In addition to these efficiency-focused models, variations such as BERT-base-cased, BERT-base-uncased, and RoBERTa reflect different choices in pretraining objectives and tokenization strategies, providing useful benchmarks for evaluating trade-offs between accuracy and computational cost. Together, these models offer a broader landscape for examining how transformer-based architectures can balance performance with resource efficiency in line with the goals of Green AI.

- **RQ2.** How do smaller and more efficient models like DistilBERT and ALBERT compare to full-sized models like BERT and RoBERTa in terms of performance and computational efficiency?

To address these questions, we fine-tuned several BERT-based models that have demonstrated strong performance across a wide range of natural language processing (NLP) tasks. Fine-tuning involves adapting a general-purpose language model that has already been pretrained on large datasets such as Wikipedia (800 million words) and BookCorpus (2.5 billion words) [7]. These pretrained models, made publicly available by their developers, serve as a foundation for more specialized tasks.

For this study, we fine-tuned the models using two labeled datasets: WELFake and ISOT Fake News. Each dataset contains examples of both fake and factual news articles, enabling the models to distinguish between the two using supervised learning [26]. During fine-tuning, the models automatically updated their internal parameters based on these examples, learning to identify statistical patterns and linguistic features commonly associated with misinformation.

After fine-tuning on the WELFake and ISOT datasets, each model was evaluated by making predictions on a held-out test set composed of news articles not seen during training. This step, known as inference, was used to assess how accurately and efficiently the models could classify new inputs.

We applied a consistent setup across all models to optimize them for fake news detection. Our goal was to evaluate how effectively each model distinguishes between fake and factual news and to determine whether transformer-based models offer a meaningful advantage over traditional machine learning approaches. Along with measuring classification accuracy, we recorded each model's runtime as a proxy for computational cost, which also reflects its potential environmental impact. This analysis helps identify models that offer a practical balance between performance and efficiency.

3 Method

3.1 Dataset Description

This study used two widely recognized public datasets: WELFake and the ISOT Fake News Dataset. Both are widely used in fake news detection research and provide a large number of labeled articles suitable for fine-tuning and evaluation. An important strength of these datasets is their relative balance between fake and factual news articles. Balanced datasets help prevent model bias, improve learning across both categories, and ensure that evaluation metrics provide an accurate reflection of model performance. This makes them well suited for training effective and reliable fake news detection models.

The WELFake dataset [30] contains 72,134 news articles labeled as either fake or factual. It was created by merging content from four existing datasets: Kaggle, McIntire, Reuters, and BuzzFeed Political. The dataset includes 37,106 fake articles and 35,028 factual

articles. Due to its size and diversity, WELFake is widely used in misinformation research. Its variety of topics and writing styles helps expose models to a broad range of linguistic patterns, improving their ability to generalize and detect fake news across different contexts.

The ISOT Fake News dataset [2] was curated by the Information Security and Object Technology (ISOT) Lab at the University of Victoria in Canada. It includes 21,417 factual articles collected from Reuters.com between January 2016 and December 2017, and 23,502 fake news articles gathered from unreliable websites flagged by PolitiFact between March 2015 and February 2018. This dataset was included to complement WELFake and to test whether findings from the model evaluations generalize across different data sources.

3.2 Data Preprocessing

Prior to fine-tuning, all text data was preprocessed to ensure format uniformity and reliable input across models. This involved converting all text entries to string format and dropping rows with null values in the "text" column. Each dataset was then randomly divided into a training set comprising 80% of the data and a testing set comprising the remaining 20%. The same data-splitting strategy was applied across all models to ensure a fair basis for comparison. Finally, the cleaned data was converted into the HuggingFace dataset format for compatibility with transformer-based models during fine-tuning.

3.3 Model Selection

From the BERT family of transformer-based models, we selected five widely used variants: BERT-base-cased, BERT-base-uncased, RoBERTa-base, DistilBERT-base-uncased, and ALBERT-base-v2. These models were chosen for their strong language modeling capabilities and varying architectural optimizations, which allow us to compare their performance under different resource constraints. By selecting these BERT-based models, we aim to explore the trade-offs between accuracy, efficiency, and adaptability across diverse text styles.

BERT-base-cased: BERT trained on cased text, distinguishing between uppercase and lowercase letters. It consists of 12 layers, 768 hidden units, and 12 attention heads, with approximately 110 million parameters [7].

BERT-base-uncased: Trained without distinguishing case; shares the same architecture as BERT-base-cased but is more generalizable for tasks where case sensitivity is not crucial [7].

RoBERTa-base: RoBERTa (Robustly Optimized BERT Approach) retains BERT's architecture but removes the Next Sentence Prediction objective and is trained on a corpus approximately ten times larger [18], often improving text classification performance [27].

DistilBERT-base-uncased: A smaller, faster version of BERT with about 60% of the parameters while retaining roughly 97% of performance [23]. Its reduced size lowers inference cost without significantly compromising accuracy.

ALBERT-base-v2: A Lite BERT (ALBERT) reduces parameters via cross-layer weight sharing and factorized embeddings [17]. Despite its compactness, ALBERT achieves competitive performance [3, 4].

Table 1: Performance comparison of traditional and hybrid models on fake news detection (WELFake dataset).

Model	Acc.	Prec.	Rec.	F1
CNN+LSTM+FastText	99.00%	99.00%	99.00%	99.00%
N-gram + TF-IDF + LSTM	96.00%	95.50%	96.20%	95.80%
Hybrid CNN+BiLSTM (Word2Vec)	97.74%	N/A	N/A	N/A
CNN–DNN Hybrid	97.32%	97.30%	97.30%	97.30%
WELFake Model (Voting)	96.73%	N/A	N/A	N/A
KNN	90.16%	N/A	N/A	N/A
SVM	96.73%	N/A	N/A	N/A
Average	**96.24%**	**97.27%**	**97.50%**	**97.37%**

3.4 Fine-Tuning

Each model was fine-tuned on the WELFake and ISOT Fake News datasets, updating weights to optimize performance on fake news detection. To ensure comparability, we applied the same configuration across all models:

Learning Rate: 2e-5, a small step size for stable, gradual learning [7].

Batch Size: 16, balancing efficiency with manageable memory use.

Epochs: 3, a standard choice for transformers that avoids overfitting [18].

Optimizer: AdamW, which supports effective learning while mitigating over-regularization [19].

Training was conducted on a high-end workstation with an NVIDIA RTX 3090 GPU and 128 GB RAM, providing sufficient capacity for fine-tuning without requiring large-scale computing clusters. This setup highlights that transformer models can be effectively fine-tuned in accessible environments, supporting reproducibility by other researchers.

4 Results

The model performance was evaluated using accuracy, precision, recall, F1 score, and runtime. Accuracy measures overall correctness. Precision reflects the proportion of items labeled as fake that were actually fake. Recall reflects the proportion of all fake items correctly identified. The F1 score combines precision and recall. Runtime assesses speed and is a proxy for computational cost.

4.1 Baseline Machine Learning Models

Traditional and hybrid models (summarized in Table 1) illustrate a range of machine learning approaches to fake news detection, commonly evaluated on the WELFake dataset. Examples include CNN+LSTM+FastText [14], CNN+BiLSTM with Word2Vec [21], TF–IDF with LSTM [16], and CNN–DNN hybrids [9]. Ensemble voting improves robustness [30], and classic models such as k-Nearest Neighbors (KNN) and Support Vector Machines (SVMs) remain competitive with appropriate features [30].

4.2 WELFake Dataset Results

Table 2 presents results on WELFake, where BERT-based models significantly outperformed traditional approaches. The average accuracy of seven traditional models was 96.24% (error 3.76%) versus 99.47% accuracy (error 0.53%) for five BERT-family models. The

Table 2: Summary of results: WELFake dataset.

Model	Acc.	Prec.	Rec.	F1
BERT-base-cased	99.52%	99.49%	99.57%	99.53%
BERT-base-uncased	99.52%	99.60%	99.47%	99.54%
RoBERTa	99.91%	99.95%	99.88%	99.92%
DistilBERT-base-uncased	98.75%	98.38%	99.20%	98.79%
ALBERT-base-v2	99.66%	99.66%	99.69%	99.68%
Average	**99.47%**	**99.42%**	**99.56%**	**99.49%**

Table 3: Summary of results: ISOT Fake News dataset.

Model	Acc.	Prec.	Rec.	F1
BERT-base-cased	99.87%	99.95%	99.80%	99.87%
BERT-base-uncased	99.80%	99.90%	99.70%	99.80%
RoBERTa	99.87%	99.85%	99.90%	99.87%
DistilBERT-base-uncased	99.78%	99.85%	99.72%	99.77%
ALBERT-base-v2	99.83%	99.85%	99.81%	99.82%
Average	**99.83%**	**99.88%**	**99.79%**	**99.83%**

odds of misclassification were approximately 7.33 times higher for traditional models than for BERT models ($p < .001$). In other words, BERT-based models reduced the likelihood of incorrect classification by 86% compared to traditional approaches. These findings demonstrate the superior performance of transformer-based models in detecting fake news and their ability to capture nuanced patterns in textual data.

In terms of runtime, BERT models did not differ significantly. Fine-tuning took about 1 hour 15 minutes on average (SD = 4 minutes), indicating gains were not due to substantially longer training. This corresponds to an estimated 750 Wh, roughly the emissions from 150 g of coal (back-of-envelope estimate). Cased vs. uncased BERT variants showed no meaningful performance gaps. RoBERTa achieved the highest accuracy, precision, recall, and F1, likely due to optimized pretraining (no Next Sentence Prediction; larger corpora).

Distilled models (DistilBERT, ALBERT) were not markedly faster to fine-tune but matched larger models' accuracy with far fewer parameters (DistilBERT ~66M across 6 layers; ALBERT ~12M with weight sharing across 12 layers), making them compelling efficiency choices.

4.3 ISOT Fake News Dataset Results

Table 3 shows that BERT-family models again outperformed traditional approaches on ISOT. The BERT-family averaged 99.83% accuracy (error 0.17%), compared to 96.24% (error 3.76%) for traditional models. Misclassification odds were 22.94 times higher for traditional models ($p < .001$), representing a 95.6% reduction using BERT-based models.

All five BERT-family models achieved consistently high accuracy, precision, recall, and F1, underscoring their robustness and generalizability across datasets. Fine-tuning required on average 19 minutes (SD $\approx$ 4 min), corresponding to 200 watt-hours of energy. DistilBERT was especially efficient, maintaining similar accuracy while producing emissions equivalent to just 20 grams of coal, highlighting the sustainability advantages of smaller transformer architectures.

Fine-tuning on ISOT averaged 19 minutes 6 seconds (SD = 4 minutes 14 seconds), estimated 200 Wh (~40 g coal). DistilBERT was especially efficient (~20 g coal equivalent) with similar accuracy. ALBERT was not faster but achieved comparable accuracy with a compact architecture.

5 Discussion

This study highlights the significant advantages of small language models in detecting fake news, particularly compared to traditional machine learning approaches. Unlike fact-checking systems that rely on external knowledge bases, our approach analyzes linguistic features. The BERT models we fine-tuned detect fake news by learning patterns in tone, structure, and word choice. Results from both the WELFake and ISOT datasets show that transformer-based models consistently outperform earlier methods. With average accuracy rates of 99.47% and 99.83%, they reduced the likelihood of misclassification by approximately 86% and 96%, underscoring the effectiveness of language-based cues for identifying misinformation.

What sets transformer models apart is their ability to capture linguistic context, enabling them to distinguish between factual and fake news with nuance that traditional models often miss. This capacity is crucial for fake news detection, where even subtle differences in wording can affect accuracy. Our findings reinforce that transformer models, especially BERT and its variants, excel in tasks requiring contextual understanding.

We also observed strong performance from smaller models, which offer faster processing and greater efficiency. Despite fewer parameters, they maintain high accuracy, showing that strong results do not always require resource-heavy architectures. DistilBERT, in particular, balances efficiency and performance, making it well-suited for limited computing environments. It also offers a more sustainable option: fine-tuning DistilBERT on ISOT produced emissions equivalent to burning just 20 grams of coal. This demonstrates that high-quality results can be achieved with far less environmental cost, pointing to a promising direction for future research.

The efficiency of these models also benefits mobile and edge computing. Edge computing processes data locally rather than on centralized servers [5]. Lightweight transformer models such as DistilBERT can run on personal devices, enabling local processing without cloud dependence [32]. This shift supports accessibility, privacy, and real-time use, allowing individuals to detect fake news directly on their devices and opening possibilities for scalable, privacy-preserving misinformation detection.

Limitations and Future Directions. There are several limitations to this study. First, we relied only on the WELFake and ISOT datasets, which, while balanced and diverse, may not capture the full range of real-world fake news. Future work should test additional datasets to strengthen generalizability.

Second, we compared our models to traditional approaches using results from prior studies rather than running them under identical conditions. Although this limits direct comparison, the breadth of prior work still provides meaningful context.

Third, runtime was used as a proxy for energy consumption. While GPU utilization was high, we did not measure power usage directly. Future studies should monitor energy consumption more systematically and also report inference speed, which was not assessed here.

Finally, our experiments required a high-end PC. To enhance accessibility, future research should test model feasibility on more common, resource-limited systems such as laptops.

These directions will help improve both the robustness and sustainability of small language models for fake news detection.

6 Conclusion

Overall, this study demonstrated the effectiveness of small language models for fake news detection, showing that they outperform traditional machine learning methods. All of the models we tested achieved high accuracy, underscoring the potential of transformer-based approaches even at smaller scales. The findings also highlight the advantages of models like DistilBERT, which offer comparable performance with significantly greater resource efficiency. Looking ahead, there is strong potential to further develop Green AI strategies that improve the sustainability of these models while continuing to enhance their accuracy and practical applicability.

References

[1] Arush Agarwal and Akhil Dixit. 2020. Fake news detection: an ensemble learning approach. In *2020 4th international conference on intelligent computing and control systems (ICICCS)*. IEEE, 1178–1183.

[2] Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *Security and Privacy* 1, 1 (2018), e9.

[3] Shafna Fitria Nur Azizah, Hasan Dwi Cahyono, Sari Widya Sihwi, and Wisnu Widiarto. 2023. Performance analysis of transformer based models (BERT, ALBERT, and RoBERTa) in fake news detection. In *2023 6th International Conference on Information and Communications Technology (ICOIACT)*. IEEE, 425–430.

[4] Hyunjin Choi, Judong Kim, Seongho Joe, and Youngjune Gwon. 2021. Evaluation of bert and albert sentence embedding performance on downstream nlp tasks. In *2020 25th International conference on pattern recognition (ICPR)*. IEEE, 5482–5487.

[5] Dylan De Beer and Machdel Matthee. 2020. Approaches to identify fake news: a systematic literature review. In *International Conference on Integrated Science*. Springer, 13–22.

[6] Alex De Vries. 2023. The growing energy footprint of artificial intelligence. *Joule* 7, 10 (2023), 2191–2194.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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 (long and short papers)*. 4171–4186.

[8] Jia Ding, Yongjun Hu, and Huiyou Chang. 2020. BERT-based mental model, a better fake news detector. In *Proceedings of the 2020 6th international conference on computing and artificial intelligence*. 396–400.

[9] Ahmed Elmasry and Nourhan Talal. 2024. Differentiating Authentic News from Fabrications using Deep Learning: A New Approach. *Systems Assessment and Engineering Management* 1 (2024), 45–53. doi:10.61356/j.saem.2024.1272

[10] Environmental Protection Agency (EPA). 2024. Greenhouse gas equivalencies calculator. https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator. Accessed: 2025-09-12.

[11] Parvathy Ganesh, Lekshmi Priya, and R Nandakumar. 2021. Fake news detection-a comparative study of advanced ensemble approaches. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 1003–1008.

[12] Cindy. Gordon. 2024. ChatGPT and generative AI innovations are creating sustainability havoc. Forbes. https://www.forbes.com/sites/cindygordon/2024/03/12/chatgpt-and-generative-ai-innovations-are-creating-sustainability-havoc/, Accessed: 2025-09-12.

[13] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2019. The future of misinformation detection: new perspectives and trends. *arXiv preprint arXiv:1909.03654* (2019).

[14] Ehtesham Hashmi, Sule Yildirim Yayilgan, Muhammad Mudassar Yamin, Subhan Ali, and Mohamed Abomhara. 2024. Advancing fake news detection: Hybrid deep learning with fasttext and explainable ai. *IEEE Access* 12 (2024), 44462–44480.

[15] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications* 80, 8 (2021), 11765–11788.

[16] Nabeela Kausar, Asghar AliKhan, and Mohsin Sattar. 2022. Towards better representation learning using hybrid deep learning model for fake news detection. *Social Network Analysis and Mining* 12, 1 (2022), 165.

[17] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[19] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[20] Mohammed Lubbad. 2023. GPT-4 parameters: Unlimited guide NLP's game-changer. https://mlubbad.medium.com/the-ultimate-guide-to-gpt-4-parameters-everything-you-need-to-know-about-nlps-game-changer-109b8767855a. Accessed: 2025-09-12.

[21] Mohamed-Amine Ouassil, Bouchaib Cherradi, Soufiane Hamida, Mouaad Errami, Oussama el Gannour, and Abdelhadi Raihani. 2022. A fake news detection system based on combination of word embedded techniques and hybrid deep learning model. *International Journal of Advanced Computer Science and Applications* 13, 10 (2022).

[22] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350* (2021).

[23] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[24] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Commun. ACM* 63, 12 (2020), 54–63.

[25] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13693–13696.

[26] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification?. In *China national conference on Chinese computational linguistics*. Springer, 194–206.

[27] Joan C Timoneda and Sebastián Vallejo Vera. 2025. BERT, RoBERTa, or DeBERTa? Comparing Performance Across Transformers Models in Political Science Text. *The Journal of Politics* 87, 1 (2025), 347–364.

[28] TRG Datacenters. 2023. AI chatbots: Energy usage of 2023's most popular chatbots (so far). https://www.trgdatacenters.com/resource/ai-chatbots-energy-usage-of-2023s-most-popular-chatbots-so-far/. Accessed: 2025-09-12.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[30] Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. 2021. WELFake: word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems* 8, 4 (2021), 881–893.

[31] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science* 359, 6380 (2018), 1146–1151.

[32] Yue Zheng, Yuhao Chen, Bin Qian, Xiufang Shi, Yuanchao Shu, and Jiming Chen. 2025. A review on edge large language models: Design, execution, and applications. *Comput. Surveys* 57, 8 (2025), 1–35.

[33] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.