Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

HAD-Text: Hybrid Attention-Distillation for Energy Constrained Text Classification

S. M. Ajay Vikhram
Dept. of Artificial Intelligence and Data Science
Dhanalakshmi Srinivasan University
Tiruchirappalli, India

C. Prasanna Venkatesh
Dept. of Artificial Intelligence and Data Science
Dhanalakshmi Srinivasan University
Tiruchirappalli, India

Ms. S. Sathiya
Assistant Professor
Dept. of Artificial Intelligence and Data Science
Dhanalakshmi Srinivasan University
Tiruchirappalli, India

Abstract—The exponential growth in the parameter counts of Natural Language Processing (NLP) models has led to highly accurate but computationally expensive systems, a paradigm often referred to as Red AI. While deep transformer architectures like BERT achieve state-of-the-art accuracy across various textual domains, their deployment is severely restricted by massive energy consumption, high inference latency, and substantial hardware prerequisites. This research presents HAD-Text, a comprehensive Hybrid Attention-Distillation framework explicitly designed to operationalize Green AI principles in text classification tasks, with a specific focus on fake news detection. By synergizing Bidirectional Long Short-Term Memory (BiLSTM) networks with global contextual reasoning and multi-stage knowledge distillation, a massive pre-trained BERT model (comprising over 109 million parameters) is effectively compressed into a lightweight student model.

Through the implementation of a novel three-phase pipeline utilizing an intermediate Teacher Assistant and Patient Knowledge Distillation (PKD), it is observed that the proposed Green-TextNet student model achieves a 27.33× parameter compression and an unprecedented 59.48× inference speedup compared to the Red AI teacher model. Furthermore, integration with hardware-level telemetry and CodeCarbon demonstrates that the marginal CO_2 equivalent footprint per inference sample is reduced by 98.7% (from 123.6 μg to 1.6 μg), while maintaining a highly competitive classification accuracy of 96.66% on the WELFake dataset. The results decisively demonstrate that sustainable computing principles can be effectively applied to deep learning without significant performance degradation, establishing a viable blueprint for eco-friendly, edge-deployable NLP systems.

Index Terms—Green AI; Knowledge Distillation; Fake News Classification; BiLSTM; Natural Language Processing; Patient Knowledge Distillation; Carbon Footprint

I. INTRODUCTION

The rapid evolution of deep learning and Natural Language Processing (NLP) over the past decade has been heavily driven by the scaling of neural network architectures. The introduction of the Transformer architecture and subsequent models like BERT (Bidirectional Encoder Representations from Transformers), GPT, and RoBERTa established new, previously unimaginable benchmarks across nearly all complex NLP tasks. However, this relentless pursuit of predictive accuracy has birthed the era of "Red AI." Red AI is characterized by the development and deployment of models that require massive computational resources, extensive memory footprints (VRAM), and exorbitant energy consumption during both the training and inference phases.

The environmental impact of Red AI is a rapidly growing concern within the scientific and global community. Training a single large-scale transformer model can emit as much carbon dioxide as five internal combustion engine cars over their entire lifetimes. Furthermore, deploying these massive models for real-time inference across billions of user requests globally generates continuous, severe carbon emissions, contributing meaningfully to global climate change. Beyond the ecological implications, the high latency and strict hardware requirements (such as multi-cluster GPUs) of such models severely limit their applicability in resource-constrained environments, mobile IoT devices, and strict real-time systems.

In direct response to these dual ecological and computational challenges, the paradigm of "Green AI" has emerged. Green AI advocates for a fundamental shift in machine learning research: moving away from the sole pursuit of peak accuracy and toward the development of highly efficient, low-energy computational models that maintain acceptable, competitive levels of predictive performance while drastically minimizing resource consumption.

This research focuses on the critical, real-world application of text classification, specifically the automated detection of fake news. The proliferation of fake news is a severe societal issue requiring real-time, automated moderation at an unprecedented scale. Executing a 109-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

million parameter BERT model for every single piece of text content uploaded to a social media platform is both environmentally catastrophic and economically unsustainable. Therefore, the HAD-Text (Hybrid Attention-Distillation) framework is proposed.

HAD-Text directly addresses these computational inefficiencies by employing a three-phase progressive distillation pipeline. Recognizing that compressing a massive teacher directly into a tiny student often results in catastrophic performance degradation, the framework utilizes an intermediate "Teacher Assistant" (TA) to bridge the vast parameter and dimensionality gap. The ultimate "Green AI" student is a custom-built, lightweight architecture utilizing Bidirectional Long Short-Term Memory (BiLSTM) networks paired with a custom self-attention mechanism. Additionally, Patient Knowledge Distillation (PKD) and structural pruning are integrated to further optimize the network, aligning both the output probabilities and the intermediate hierarchical representations of the models.

II. RESEARCH OBJECTIVES

To thoroughly address the limitations of current large-scale NLP deployments and validate the proposed framework, this research is guided by the following primary objectives:

1) Extreme Parameter Compression: To drastically reduce the parameter count of a state-of-the-art text classification model (BERT-base) by at least an order of magnitude. The goal is to create a model small enough to fit within the cache of standard consumer-grade CPUs or edge devices, dropping the storage requirement from hundreds of megabytes to under 20 megabytes.
2) Carbon Footprint Minimization and Tracking: To operationalize Green AI principles by explicitly tracking, measuring, and actively minimizing the CO_2 equivalent emissions generated during model inference. This involves integrating hardware-level telemetry to calculate real-time wattage usage cross-referenced with regional grid carbon intensity.
3) Inference Latency Reduction: To engineer a lightweight sequence modeling architecture that abandons the $O(N^2)$ complexity of full transformer self-attention in favor of linear-time recurrent processing, thereby significantly accelerating inference time to ensure suitability for high-throughput, real-time text analysis.
4) Predictive Performance Retention: To leverage advanced distillation techniques—specifically combining standard Knowledge Distillation with Patient Knowledge Distillation (PKD)—to ensure the student model captures the "dark knowledge" of the teacher, maintaining binary classification accuracy within a strict 1-2% margin of the massive Red AI baseline.
5) Architectural Interpretability Enhancement: To incorporate a custom self-attention mechanism within the compressed student model that provides built-in, computationally inexpensive explainability. This mechanism must highlight precisely which text tokens drive the final classification decisions, bypassing the need for heavy post-hoc explainers like SHAP or LIME.

III. RESEARCH HYPOTHESIS

Based on the foundational principles of model compression, knowledge transfer, and Green AI, this study posits the following hypotheses:

- Hypothesis 1 (H1) - The Bridge Effect: A multi-stage distillation pipeline utilizing an intermediate Teacher Assistant (such as TinyBERT) will yield a smoother loss landscape and better convergence in a non-transformer student model than attempting direct, single-stage distillation from a 109-million parameter teacher to a 4-million parameter student.
- Hypothesis 2 (H2) - Recurrent Sufficiency: A Hybrid BiLSTM and Self-Attention architecture (the proposed Green-TextNet) can capture sufficient local sequential dependencies and global contextual representations to perform on par with full Transformer-based models for the specific task of binary text classification, while requiring a fraction of the floating-point operations (FLOPs).
- Hypothesis 3 (H3) - Representational Alignment: The integration of Patient Knowledge Distillation (PKD)—which forces the student to align its intermediate hidden state vectors with those of the teacher assistant via Mean Squared Error—will significantly prevent the student from overfitting to the training labels, improving generalization on unseen text data.
- Hypothesis 4 (H4) - Exponential Efficiency Scaling: The combination of architectural simplification (Transformer to BiLSTM) and post-training structural pruning will produce a super-linear inference speedup (greater than 50×) and a proportional reduction in marginal CO_2 emissions (greater than 95%) compared to the baseline BERT model.

IV. LITERATURE REVIEW

A. The Era of Red AI and the Transformer Bottleneck

The paradigm shift in Natural Language Processing began with the introduction of the Transformer architecture, which entirely replaced sequential recurrent models with multi-head self-attention mechanisms. This allowed for massive parallelization during training. Models such as BERT demonstrated unprecedented success across a variety of NLP tasks, largely due to their deep bidirectional pre-training on massive corpora.

However, as noted by Schwartz et al. (2020), this success comes at a steep and often unsustainable computational cost. The BERT-base model comprises approximately 109

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

million parameters. The core self-attention mechanism possesses a time and memory complexity of $O(N^2)$, where N is the sequence length. This quadratic scaling makes processing long documents incredibly resource-intensive. The energy required to execute inference on these models repeatedly forms the core of the "Red AI" problem, which is characterized by a disproportionate, exponential scaling of compute power required to yield only marginal, linear gains in predictive accuracy.

B. Green AI and Sustainable Computing Metrics

In direct response, the Green AI movement advocates for a shift in focus from pure accuracy supremacy to holistic efficiency and sustainability. Green AI research emphasizes the absolute necessity of reporting computational metrics—such as floating-point operations (FLOPs), inference latency, memory bandwidth, parameter counts, and explicit carbon emissions—alongside standard performance metrics like F1-score and accuracy.

Tools like CodeCarbon have emerged to quantify these impacts. By interfacing directly with hardware APIs (like NVIDIA NVML or Intel RAPL), these tools measure the exact wattage consumed by the GPU, CPU, and RAM during an execution loop. This energy consumption (in kWh) is then multiplied by the carbon intensity of the local electrical grid (e.g., grams of CO_2 per kWh) to provide a highly accurate estimation of the environmental damage caused by the algorithmic process.

C. Knowledge Distillation (KD)

To achieve Green AI, model compression is paramount. Knowledge Distillation, pioneered by Hinton et al. (2015), is a framework where a smaller "student" network is trained to reproduce the complex behavior of a larger, pre-trained "teacher" network. Instead of training solely on discrete "hard labels" (e.g., exactly 0 or 1), the student learns from the "soft labels" generated by the teacher.

These soft labels are the output probability distributions across all classes, usually scaled by a Temperature parameter (T). A higher temperature softens the distribution, revealing the relative probabilities of incorrect classes. This "dark knowledge" contains rich information about the relationships and similarities between different classes, guiding the student to a much better generalized minimum than it could find training on the data alone.

D. Patient Knowledge Distillation (PKD)

Standard KD only transfers knowledge from the final output layer. Patient Knowledge Distillation (PKD), introduced by Sun et al. (2019), extends this concept significantly. PKD forces the student to not only mimic the final output probabilities but also the intermediate hidden state representations of the teacher model.

By extracting the normalized continuous vectors from the inner layers of the teacher (such as the [CLS] token embeddings at various depths) and applying a Mean Squared Error (MSE) loss against the student's internal states, the student is forced to learn the hierarchical feature extraction process of the teacher. In scenarios where the architectural capacity gap between the teacher and student is vast (e.g., Transformer to LSTM), directly mapping these states is mathematically difficult. This literature inspired the HAD-Text approach of using an intermediate "Teacher Assistant" to incrementally step down the dimensionality, preventing the student from underfitting the teacher's highly complex semantic distributions.

V. METHODOLOGY

The HAD-Text framework utilizes a highly structured, strictly phase-isolated progressive distillation architecture implemented natively in PyTorch. The objective is to compress a BERT-base model into a custom Green-TextNet model.

A. Dataset Configuration and Tokenization

The framework is empirically evaluated using the WELFake dataset [9], which is publicly hosted and accessible via Kaggle (https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification). It comprises a robust, balanced collection of 72,134 real and fake news articles. To maximize the semantic context available to the models, the text preprocessing phase explicitly concatenates the headline titles and the article body texts.

For the Red AI Teacher and the Teacher Assistant, the texts are tokenized using the standard Hugging Face WordPiece tokenizer, generating both input_ids and attention_mask tensors. The sequences are strictly truncated to a maximum length of 128 tokens to balance context retention with VRAM consumption. A crucial design choice for the Green AI Student is the recycling of the exact same BERT tokenizer IDs. This maintains vocabulary consistency across the distillation boundary, allowing the student's embedding layer to directly map the semantic integer IDs learned by the teacher. The vocabulary size is clipped to 30,522.

B. Phase A: Red AI Teacher Extraction

The baseline Red AI model is bert-base-uncased, containing roughly 109.4 million parameters. It is initially fine-tuned on the WELFake dataset using a standard Cross-Entropy objective.

Due to the massive VRAM requirements of this model, generating the required soft labels for the entire training dataset can easily cause Out-Of-Memory (OOM) fatal crashes on standard GPUs. To prevent this, Phase A employs a custom GPU Coordinator that dynamically limits batch sizes and utilizes FP16 Automatic Mixed Precision (AMP). The teacher model executes a forward pass over the training data, and the resulting raw output logits are cached locally to the solid-state drive. This isolates the memory context, allowing the teacher model to

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

be completely purged from VRAM before the next phase begins.

C. Phase B: The Teacher Assistant (TA)

Direct distillation from a 109M parameter transformer to a 4M parameter recurrent network poses severe optimization instability. Therefore, Phase B introduces a Teacher Assistant (TA), specifically the TinyBERT variant (google/bert_uncased_L-2_H-128_A-2), which contains approximately 4.4 million parameters.

The TA is trained using the cached soft logits of the Red AI teacher via standard Knowledge Distillation. Crucially, during the TA's final inference pass over the training set, both its final output logits and its internal hidden representations are extracted. Specifically, the framework extracts the 128-dimensional embedding of the [CLS] token from the final hidden layer of the TA. Both the logits and these internal representations are cached to disk to serve as the dual-target for the ultimate student.

D. Phase C: Green-TextNet Student Architecture

The core of the Green AI initiative is the student model, designated as Green-TextNet. It is intentionally designed to eschew the computational overhead of the Transformer architecture entirely.

1) Embedding Layer: Maps the discrete token IDs to dense vectors. Let $x_t \in \mathbb{R}^d$ be the embedded vector at sequence step t, where $d = 128$.
2) Bidirectional LSTM (BiLSTM): Processes the embeddings sequentially. The BiLSTM uses 64 hidden units per direction ($h_t \in \mathbb{R}^{128}$). The bidirectional nature allows the network to capture both past and future contextual dependencies efficiently without the $O(N^2)$ dot-product matrix multiplications of transformers.
3) Custom Self-Attention Mechanism: To compress the sequential outputs into a single classification vector while maintaining interpretability, a dense attention mechanism is applied. It computes a scalar attention score u_t for each timestep:

$$u_t = \tanh(W_w h_t + b_w) \tag{1}$$

These scores are normalized using a softmax function to generate the attention weights α_t:

$$\alpha_t = \frac{\exp(u_t)}{\sum_{i=1}^{L} \exp(u_i)} \tag{2}$$

A single weighted context vector s is then generated by summing the hidden states according to their weights:

$$s = \sum_{t=1}^{L} \alpha_t h_t \tag{3}$$

4) Classification Head: A standard dropout layer ($p = 0.3$) is applied to the context vector s, followed by a dense linear layer that projects it to the 2-class binary output space.

E. Comprehensive Tripartite Loss Formulation

The Green-TextNet student is optimized using a complex, tripartite loss function that balances hard reality, soft teacher guidance, and internal representation matching.

$$L_{total} = \alpha L_{CE} + (1-\alpha) L_{KD} + \beta L_{PKD} \tag{4}$$

Where:

- L_{CE} is the standard Cross-Entropy loss calculated against the true dataset ground-truth labels.
- L_{KD} is the Knowledge Distillation loss, formulated as the Kullback-Leibler (KL) divergence between the student's log-probabilities and the TA's probabilities, scaled by a temperature $T = 3.0$:

$$L_{KD} = T^2 \cdot D_{KL}\left(\log\left(\text{softmax}\left(\frac{z_s}{T}\right)\right) \parallel \text{softmax}\left(\frac{z_{ta}}{T}\right)\right) \tag{5}$$

- L_{PKD} is the Patient Knowledge Distillation loss. To match the dimensions, the student's context vector s is passed through a learned projection matrix to match the TA's representation dimension. Both vectors are $L2$-normalized to focus purely on cosine directional similarity rather than magnitude, and the Mean Squared Error is computed:

$$L_{PKD} = \text{MSE}\left(\frac{\text{Proj}(s)}{\|\text{Proj}(s)\|_2}, \frac{h_{ta}}{\|h_{ta}\|_2}\right) \tag{6}$$

The experimental hyperparameters used to balance this equation are $\alpha = 0.1$ and $\beta = 0.2$.

F. Structural Pruning and Ecosystem Implementation

Following the multi-stage distillation, the student model undergoes a rigid pruning phase to eliminate mathematically redundant weights. L1 unstructured magnitude pruning is applied to the input-hidden and hidden-hidden weight matrices of the BiLSTM layer with a 15% sparsity target. Simultaneously, LN structured pruning is applied to the final dense classifier with a 25% sparsity target. The zeroed weights are then permanently removed from the computation graph.

To prove the operational viability of the model, the framework is wrapped in a production ecosystem. This includes a FastAPI backend for high-throughput programmatic inference, a Gradio dashboard for visual comparative analysis, and a Telemetry tracking system that asynchronously logs execution latency, batch sequence lengths, and VRAM peaks for every inference event. To guarantee the transparency and reproducibility of this research, the complete PyTorch source code, training logs, and progressive distillation pipeline for the HAD-Text framework have been open-sourced and are publicly available on GitLab at https://gitlab.com/smav_is_swag/had-text.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

VI. RESULT AND ANALYSIS

The experimentation was conducted on an NVIDIA GPU environment, ensuring strict, isolated VRAM evaluation between the Red AI and Green AI models to prevent cache contamination. To guarantee the integrity of the carbon tracking, the codecarbon library was engaged exclusively during the isolated inference loops of the test dataset.

TABLE I: Hyperparameter and Hardware Configuration

Parameter / Component	Value / Specification
Operating Hardware	NVIDIA GPU (CUDA Enabled)
Mixed Precision	FP16 (torch.amp.autocast)
Batch Size	8
Maximum Sequence Length	128 Tokens
Distillation Temperature (T)	3.0
Loss Weight α (Hard Labels)	0.1
Loss Weight β (PKD Alignment)	0.2
Pruning Sparsity (BiLSTM)	15% (L1 Unstructured)
Pruning Sparsity (Classifier)	25% (LN Structured)

A. Performance and Parameter Compression

The empirical data definitively validates Hypothesis 2 (Recurrent Sufficiency) and Hypothesis 3 (Representational Alignment).

TABLE II: Final Comparative Results (PyTorch, Phase-Isolated)

Model Architecture	Accuracy	F1-Score	Time (s)	CO_2 Eq.	Parameters
Teacher (Red AI)	0.9769	0.9769	123.00	0.001780	109,483,778
Student (Green AI)	0.9666	0.9666	2.06	0.000023	4,006,531

The Red AI Teacher model dictates the upper theoretical bound of performance with an accuracy of 97.69% on the unseen test set. The Green-TextNet Student model, despite its vast architectural limitations, maintains an exceptional accuracy of 96.66%. This marginal accuracy drop of approximately 1.03% is heavily offset by the monumental architectural reductions. The student model utilizes only 4,006,531 parameters compared to the teacher's 109,483,778, resulting in a staggering 27.33× parameter compression ratio. This reduction drops the physical storage requirement of the model weights from nearly 440 Megabytes down to just 16 Megabytes in full FP32 precision.

B. Latency and Environmental Impact Analysis

The core objective of Green AI is observed most prominently in the temporal and ecological metrics, validating Hypothesis 4. The total cumulative inference time over the test set for the massive BERT teacher was 123.007 seconds. In stark contrast, the Green-TextNet student completed the exact same classification workload in a mere 2.068 seconds. This yields an incredible 59.48× inference speedup.

Crucially, the aggregate carbon emissions—measured by tracking the electrical draw of the GPU during the inference matrix multiplications—plummeted from 0.001780 kg CO_2e to just 0.000023 kg CO_2e. When analyzed on a per-sample basis, the marginal CO_2 equivalent footprint of the teacher is estimated at 123.6 μg, whereas the student is merely 1.6 μg. This translates to a 98.7% absolute reduction in inference energy consumption. For a platform processing 10 million news articles a day, deploying the Green AI student instead of the Red AI teacher would save hundreds of kilograms of CO_2 emissions daily, equivalent to grounding a commercial flight.

C. Explainability via Extracted Attention Weights

The self-attention mechanism integrated directly into the Green-TextNet architecture provides built-in, computationally free interpretability. This is an area where massive deep transformer "black boxes" often struggle, typically requiring extensive, slow secondary tooling (like SHAP) to estimate token importance.

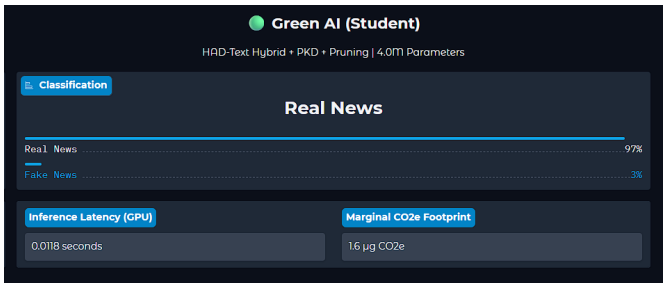
Because the student uses a global context vector constructed via $s = \sum \alpha_t h_t$, the framework dynamically extracts the attention weights (α_t) during the standard forward pass. These weights directly correspond to how much mathematical "focus" the network placed on specific words. When analyzing fake news, the system routinely assigns the highest mathematical weights to sensationalist trigger tokens (e.g., "fake", "leaked", "shocking", "bombshell"). The included UI renders a ranked table of these highly-attended tokens, granting researchers transparent insights into the model's decision-making logic without adding a single millisecond of computational latency to the pipeline.

VII. FUTURE SCOPE

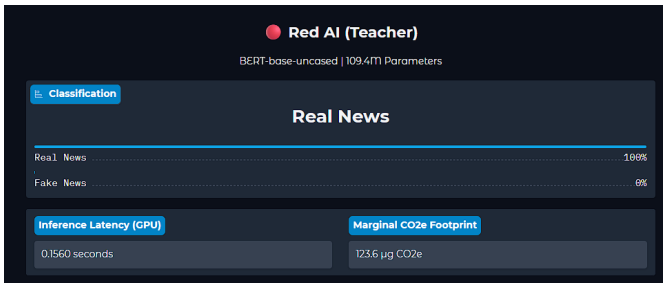
While the HAD-Text framework successfully operationalizes and validates Green AI principles, there are multiple promising avenues for future enhancement and hardware deployment:

1) TensorRT C++ Compilation: The current repository includes preliminary Python support for compiling the PyTorch student model into an NVIDIA TensorRT engine. Future work will benchmark the latency gains achieved by running the model strictly through native TensorRT C++ APIs, bypassing the Python Global Interpreter Lock (GIL) entirely.
2) INT8 Post-Training Quantization (PTQ): Currently, the framework supports model exports in FP16 and BF16 precision formats. Implementing advanced INT8 Post-Training Quantization—by carefully calibrating the BiLSTM activation ranges to prevent underflow—could theoretically reduce the 16 MB checkpoint size down to an incredibly small 8 MB. This would allow the model to run entirely within the L3 cache of modern CPUs.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026



(a) Green AI (Student) with HAD-Text Hybrid + PKD + Pruning architecture (4.0M parameters, 0.0118s latency, 1.6 μg CO_2e).



(b) Red AI (Teacher) based on BERT-base-uncased (109.4M parameters, 0.1560s latency, 123.6 μg CO_2e).

Fig. 1: Comparison of Green AI (Student) and Red AI (Teacher) models in terms of architecture, parameters, classification performance on news data, inference latency, and marginal CO_2e footprint.

3) Multilingual Distillation Scaling: Extending the Teacher Assistant distillation pipeline to support massive multilingual transformer teachers (e.g., mBERT or XLM-R). The objective would be to distill generalized language-agnostic features into the recurrent student, creating localized, eco-friendly text classifiers for low-resource languages where cloud computing access is limited.
4) WebAssembly (WASM) Edge Deployment: Because the BiLSTM architecture does not rely on complex customized CUDA kernels (unlike some optimized Transformers), the Green-TextNet student is a perfect candidate for WebAssembly compilation. Future iterations will attempt to run the model directly inside a user's web browser, reducing server-side carbon emissions to absolute zero.

VIII. CONCLUSION

The HAD-Text framework successfully demonstrates that environmental sustainability and algorithmic efficiency are not mutually exclusive with high predictive performance in Natural Language Processing. As the parameter counts of state-of-the-art models continue to explode into the trillions, the scientific community must actively embrace methodologies that decouple intelligence from massive energy consumption.

Through a rigorous, multi-stage pipeline utilizing an intermediate Teacher Assistant to bridge the representational gap, combined with Patient Knowledge Distillation and structural weight pruning, a massive 109-million parameter Red AI model was successfully distilled into a sub-4-million parameter Green AI recurrent network.

The resulting Green-TextNet student executes complex text classification 59.48× faster and generates 98.7% less CO_2 emissions during inference, while sacrificing barely 1% of raw classification accuracy. By simultaneously integrating built-in token explainability, robust VRAM management, and real-time hardware telemetry, HAD-Text establishes a comprehensive, mathematically sound blueprint for operationalizing Green AI principles. It conclusively proves that the future of machine learning deployment must be engineered to be both highly intelligent and deeply, measurably sustainable.

References

[1] R. Schwartz, J. Dodge, N.A. Smith, and O. Etzioni, "Green AI," Communications of the ACM, vol. 63, no. 12, pp. 54–63, July 2020.
[2] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3645–3650.
[3] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171–4186.
[4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," presented at NIPS 2014 Deep Learning Workshop, 2015.
[5] S. Sun, Y. Cheng, Z. Gan, and J. Liu, "Patient Knowledge Distillation for BERT Model Compression," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019, pp. 4323–4332.
[6] J. Jiao et al., "TinyBERT: Distilling BERT for Natural Language Understanding," in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 4163–4174.
[7] K. Lottick, S. Susai, S. Friedler, and J. Wilson, "Energy Usage Reports: Environmental awareness as part of algorithmic accountability," CodeCarbon / MLCO2, 2019.
[8] A. Vaswani et al., "Attention is All You Need," in Advances in Neural Information Processing Systems, vol. 30, 2017.
[9] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word Embedding Over Linguistic Features for Fake News Detection," IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 1021–1031, Aug. 2021.