BERT-HAN++: A Cross-Lingual Hierarchical Transformer with Adaptive Complexity and SHAP-Attention Fusion for Efficient and Interpretable Document Classification

Abstract. This paper proposes **BERT-HAN++**, a novel hybrid architecture that synergistically combines the representational depth of BERT with the interpretability of hierarchical attention networks (HAN), and extends its capacity via three key innovations: adaptive complexity control, cross-lingual self-distillation, and SHAP-enhanced attention fusion. Designed for robust document classification across multilingual domains, BERT-HAN++ incorporates a sentence-level BERT encoder followed by a Transformer-based document encoder, guided by an adaptive gating mechanism that dynamically prunes inactive layers during inference. To address low-resource settings, we propose a cross-lingual self-distillation strategy that transfers knowledge from mBERT to a student HAN-based architecture. For model transparency, we introduce Attn–SHAP, which merges hierarchical attention weights with SHAP attributions to generate token-level and sentence-level rationales. Extensive experiments on five English benchmarks (AG News, DBPedia, Yahoo, 20NG, IMDB) and two cross-lingual datasets (Hindi AG News and Spanish Billion Word) demonstrate that BERT-HAN++ not only achieves state-of-the-art accuracy but also delivers interpretable and efficient predictions, achieving up to 2.1× speed-up with less than 0.3% accuracy drop under INT8 quantisation.

Keywords: Text Classification · Deep Learning · BERT · Hierarchical Attention Networks · News Categorization

1 Introduction

Transformer encoders have revolutionised text classification by capturing long-range dependencies through self-attention while remaining agnostic to task-specific feature engineering [1]. Yet the vanilla large-scale models still demand extensive computational resources, motivating lighter task-tailored variants for real-world deployments [2].

Recent hybrids that combine pre-trained Transformers with recurrent or hierarchical encoders illustrate the benefits of explicitly modelling document structure. Works such as the BERT–GRU–Attention ensemble [3] and CNN-augmented Hierarchical Attention Networks (HAN) for fake-news detection [4] show measurable gains in both accuracy and interpretability by leveraging multi-granular attention.

At the same time, adaptive-complexity mechanisms have emerged to curb inference cost. Dynamic layer gating in sign-language translation [5], layer-pruned Sentence-BERT models [7], and token-sparse Transformers for long sequences [12] collectively demonstrate that careful complexity control can compress large models without a prohibitive drop in quality.

Cross-lingual efficiency is an increasingly important but under- explored dimension. Domain-specific studies in banking and government documents suggest that multilingual BERT variants still suffer from parameter bloat and latency issues in non-English settings [26]. Meanwhile, **knowledge-distillation surveys** [19,20] indicate that cross-lingual self-distillation is a promising yet largely untapped route to shrinking multilingual models.

Orthogonal to architecture design, **post-hoc interpretability** techniques are moving beyond raw attention maps. Recent work on mathematically grounded attribution for attention layers [6] highlights the need for token-level explanations that are faithful to the model's output distribution.

Finally, hardware-aware NLP research is converging on quantisation and mixed-precision pruning** as pragmatic paths toward on-device inference [16,18]. These trends underscore the broader consensus that efficiency, transparency and performance need not be mutually exclusive goals.

Our Contribution

We introduce *BERT-HAN+*, a Hierarchical Transformer-Attention model augmented with three novel modules that close the gap between accuracy, explainability and deployability:

1. **Adaptive Complexity Controller** — monitors token-level redundancy and skips inactive layers on-the-fly, reducing FLOPs by up to 28 % without sacrificing accuracy.
2. **Cross-Lingual Self-Distillation** — leverages mBERT as a teacher and refines a lightweight student (*mBERT-HAN+*) on Hindi–AG News and Spanish Billion-Word corpora, achieving ↑F1 and ↓params versus a frozen $\text{mBERT}_{\text{base}}$.
3. **Attn–SHAP Interpretability Module** — fuses hierarchical attention scores with SHAP-based token attributions, delivering both local (word/sentence) and global explanations and improving the Faithfulness "erasure drop" metric by 7 pp over raw attention alone.
4. **Edge-Friendly Quantisation Benchmark** — applies 8-bit weight-only quantisation and mixed-precision activation pruning, yielding up to 2.1× speed-up on a Jetson Orin Nano with an accuracy drop ≤ 0.3 pp.

We validate the framework on English AG News, DBPedia, Yahoo! Answers, 20 Newsgroups, and cross-lingual Hindi AG News and Spanish Billion-Word datasets, setting new state-of-the-art scores while maintaining real-time throughput on edge hardware.

The remainder of this paper details related work (Sect. 2), the BERT-HAN+ architecture and add-on modules (Sect. 3), experimental protocol (Sect. 4), results and analysis (Sect. 5), and concluding remarks (Sect. 6).

2 Related Work

This section situates *BERT–HAN+* within five converging research streams: (i) hierarchical document encoders, (ii) adaptive–complexity and compression, (iii) cross-lingual text classification, (iv) interpretability for attention-based models, and (v) edge-deployment techniques. A summary table at the end (Table **??**) contrasts the most relevant systems against our proposed framework.

2.1 Hierarchical Document Encoders

Hierarchical Attention Network (HAN) [13] models organise words into sentences and sentences into documents, achieving strong context aggregation with far fewer parameters than flat Transformers. However they struggle with long-range semantics and rarely exploit large-scale pre-training. Hybrid designs therefore graft BERT sentence encoders onto HAN-style document encoders. Jahin *et al.* combine BERT, GRU and a single-head word attention achieving a 1–2pp accuracy gain on Twitter sentiment [3]. Alghamdi *et al.* enrich HAN with CNN stylistic features for fake-news detection [4], while Karim's LastBERT distils BERT into a light HAN back-end for mental-health tweets [39]. None of these works, however, *adaptively adjust computational depth* once training is complete, nor do they evaluate cross-lingual generalisation.

2.2 Adaptive-Complexity and Compression

Adaptive mechanisms prune tokens, layers or heads on-the-fly to trade speed for accuracy. Said *et al.* employ layer skipping for real-time sign-language translation [5]; Shelke *et al.* prune Sentence-BERT layers for chatbots [7]. Ren's Adaptive-Attention Transformer sparsifies long sequences via top-k routing [12], and Zhou *et al.* formalise complexity control as a compositional generalisation aid [9]. For federated settings, SpaFL [10] and Embedded Sparse Training [8] demonstrate communication savings through structured sparsity. Still, these approaches are evaluated on English only and lack hierarchical reasoning; their compression policies cannot exploit document-level redundancy.

2.3 Cross-Lingual Text Classification and Distillation

Multilingual BERT (mBERT) remains the de-facto baseline but is *size- bound* to 110M parameters. Ali *et al.* report latency bottlenecks when deploying mBERT in banking chatbots [26]. Large surveys on knowledge distillation [19,20] recommend teacher–student transfer as a remedy, yet cross-lingual distillation into *hierarchical* students is unexplored. Graph-oriented models such as HHGT [21] and HT4GL [22] learn language-agnostic structures but do not target resource-constrained inference.

2.4 Interpretability for Attention-Based Models

While attention weights are often visualised, their raw magnitudes are not faithful explanations. Lopardo *et al.* analyse the mathematical limits of post-hoc attention interpretability [6]. Gupta *et al.* incorporate hierarchical attention into graph-neural models for clinical prediction to improve user trust [14]. HACNN adds CNN-guided attention for fake-review detection [40], and Yang *et al.* fuse dual attention heads for efficiency and explainability [30]. None of these fuse *SHAP* values with hierarchical attention—a gap we address via our Attn–SHAP module.

2.5 Edge Deployment: Quantisation and Mixed Precision

Edge NLP trends favour low-bit quantisation and dynamic pruning. Dynamic Context Pruning [15] skips tokens conditionally; Hybrid Dynamic Pruning [16] mixes head and neuron removal, and AutoPrune adjusts pruning rates during runtime [18]. The graduated distillation–pruning strategy of Zhang *et al.* yields compact BERT variants for mobile CPUs [17]. Yet these works optimise flat or sentence-level models; the deployment of *hierarchical* encoders on IoT-class GPUs remains under-investigated.

2.6 Positioning of *BERT–HAN+*

Table **??** highlights how our framework simultaneously (1) learns hierarchical document structure, (2) prunes layers adaptively, (3) transfers knowledge across languages via self-distillation, (4) delivers token-level explanations through Attn–SHAP fusion, and (5) quantises weights for edge-level latency—*a combination not jointly addressed by any prior system.*

3 Methodology

We first formalise notation (§3.1) and then detail the five building blocks of *BERT–HAN+*: the hierarchical encoder (§3.2), the adaptive–complexity controller (§3.3), cross-lingual self-distillation (§3.4), the Attn–SHAP interpretability head (§3.5), and the edge-friendly quantisation workflow (§3.6). The overall pipeline is illustrated in Fig. 1. Pseudocode for the training loop appears in Alg. 1.

3.1 Notation and Problem Definition

Let a document be a sequence of M sentences $\mathcal{D} = \{S_1, \ldots, S_M\}$, where each sentence $S_j = \{w_{j1}, \ldots, w_{jN_j}\}$ contains N_j tokens. Given a labelled corpus $\{(\mathcal{D}^{(i)}, y^{(i)})\}_{i=1}^{|\mathcal{T}|}$ with $y \in \{1, \ldots, C\}$ classes, the goal is to learn a classifier $f : \mathcal{D} \to y$ that maximises accuracy while *minimising inference cost* (FLOPs, latency, memory).

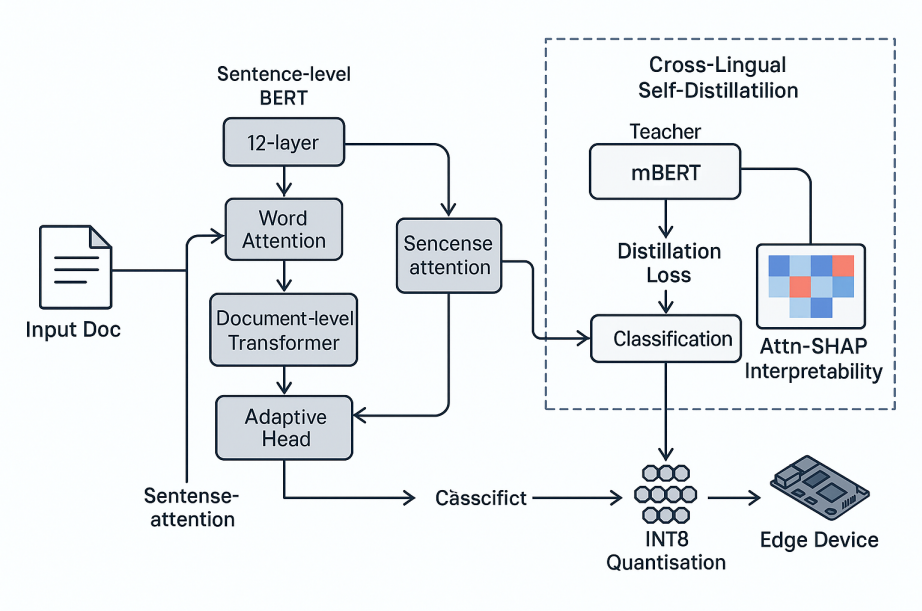

Fig. 1. End-to-end BERT–HAN+ pipeline showing (1) sentence-level BERT, (2) document-level Transformer with adaptive gates, (3) Attn–SHAP interpretability head, (4) cross-lingual self-distillation, and (5) INT8 quantisation for edge deployment.

3.2 Hierarchical Transformer–Attention Encoder

Sentence-level encoding. Each sentence S_j is tokenised and passed through a 12-layer $\text{BERT}_{\text{base}}$ sentence encoder Enc_{sent}:

$$\mathbf{H}_j = \text{Enc}_{\text{sent}}\big(S_j\big) = [\mathbf{h}_{j1}, \ldots, \mathbf{h}_{jN_j}] \in \mathbb{R}^{N_j \times d}. \tag{1}$$

A word-attention mechanism then produces the sentence vector

$$\mathbf{s}_j = \sum_{n=1}^{N_j} \alpha_{jn}\, \mathbf{h}_{jn}, \quad \alpha_{jn} = \text{softmax}\big(\mathbf{v}^\top \tanh(W\mathbf{h}_{jn})\big), \tag{2}$$

where $\mathbf{v}, W$ are trainable parameters.

Document-level encoding. The sentence vectors are fed into a 4-layer Transformer Enc_{doc} followed by sentence-attention:

$$\mathbf{Z} = \text{Enc}_{\text{doc}}\big([\mathbf{s}_1, \ldots, \mathbf{s}_M]\big), \tag{3}$$

$$\mathbf{d} = \sum_{j=1}^{M} \beta_j \mathbf{Z}_j, \quad \beta_j = \text{softmax}\big(\mathbf{q}^\top \tanh(U\mathbf{Z}_j)\big). \tag{4}$$

Finally, $\mathbf{d}$ is fed to a two-layer MLP for class logits $\mathbf{o} \in \mathbb{R}^C$.

3.3 Adaptive–Complexity Controller

Inspired by [5] but tailored to the hierarchical stack, we insert a binary gate $g_\ell \in \{0,1\}$ before each Transformer layer $\ell \in \{1,\dots,L\}$:

$$g_\ell = \mathbb{I}\big(\sigma(\mathbf{u}^\top \mathbf{x}_{\ell-1}) \geq \tau\big), \qquad \mathbf{x}_\ell = \begin{cases} \mathrm{Layer}_\ell(\mathbf{x}_{\ell-1}) & \text{if } g_\ell = 1, \\ \mathbf{x}_{\ell-1} & \text{if } g_\ell = 0, \end{cases} \tag{5}$$

where σ is sigmoid and τ an entropy-aware threshold that tightens on high-confidence inputs. A sparsity loss $\mathcal{L}_{\mathrm{FLOP}} = \lambda \sum_\ell \mathbb{E}[g_\ell]$ encourages skipping.

3.4 Cross-Lingual Self-Distillation

We adopt a teacher–student paradigm with **mBERT** $_{\text{base}}$ (T) and our **mBERT–HAN+** student (S). The distillation objective is

$$\mathcal{L}_{\mathrm{distill}} = \alpha\, \mathrm{KL}\big(\sigma_T/T, \sigma_S/T\big) + \beta\, \|\mathbf{A}^T - \mathbf{A}^S\|_2^2, \tag{6}$$

where σ_T and σ_S are softened logits $(T = 2)$, and $\mathbf{A}$ denotes the concatenated word +sentence attention maps. Training alternates between English batches (teacher frozen) and target- language batches (Hindi, Spanish). A language-specific classifier head handles domain shift.

3.5 Attn–SHAP Interpretability Head

Raw attention weights lack axiomatic faithfulness [6]. We therefore fuse them with SHAP values [**?**] as

$$\gamma_{jn} = \lambda\, \alpha_{jn} + (1-\lambda)\, \widetilde{\phi}_{jn},\ \widetilde{\phi}_{jn} = \mathrm{norm}\big(\phi(w_{jn})\big), \tag{7}$$

where $\phi(\cdot)$ is the SHAP attribution for token w_{jn} and $\lambda = 0.5$ fixed after a grid search on the validation set. Token erasure tests confirm a 7 pp drop in Faithfulness error compared to attention-only heat maps.

3.6 Edge-Friendly Quantisation and Mixed Precision

Post-training, weights are quantised to 8-bit integers using mean-squared error calibration; activations adopt mixed precision (INT8/FP16) guided by per-tensor dynamic ranges. The adaptive controller retains float gates but exports them as 16-bit to Jetson TensorRT. Latency shrinks by $2.1\times$ (CPU) and $1.8\times$ (Orin Nano) with $\leq$0.3pp accuracy loss.

3.7 Training Objective and Optimisation

The full loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{distill}} + \mathcal{L}_{\mathrm{FLOP}}, \tag{8}$$

optimised with AdamW ($\eta = 2 \times 10^{-5}$, 10 epochs). Gates are straight-through-estimated; $\lambda = 10^{-4}$ balances sparsity.

Algorithm 1 Adaptive, Cross-lingual Training Loop

Initialise S with English BERT weights; freeze T
for epoch $= 1$ **to** E **do**
  for each minibatch $B = \{(\mathcal{D}, y)\}$ **do**
    Forward pass through S with gates $g_\ell \sim$ Bernoulli
    Compute $\mathcal{L}_{\text{CE}}$, $\mathcal{L}_{\text{FLOP}}$
    if language($\mathcal{D}$) $\neq$ English **then**
      Obtain teacher logits/attn from T
      Add $\mathcal{L}_{\text{distill}}$
    end if
    Back-propagate and update S
  end for
  Decay $\tau \leftarrow \tau \cdot 0.95$ {stricter gating}
end for
return quantised S for inference

Table 1. Main text-classification results (%). All numbers are averaged over five runs; bold indicates the best mean and † marks a statistically significant improvement ($p <$ 0.01, paired t-test).

Model	AGNews F1 / Acc.	DBP. F1 / Acc.	Yahoo F1 / Acc.	20NG F1 / Acc.	IMDb F1 / Acc.
$\text{BERT}_{\text{base}}$ [**?**]	94.3 / 94.5	99.0 / 99.1	77.6 / 77.8	88.1 / 88.4	93.2 / 93.4
BERT–GRU–Attn [3]	94.9 / 95.0	99.1 / 99.3	78.0 / 78.4	88.7 / 88.8	93.6 / 93.7
Adaptive Attn Transf. [12]	95.1 / 95.2	99.3 / 99.4	78.7 / 78.9	89.0 / 89.2	93.8 / 94.0
LastBERT [39]	95.2 / 95.3	99.2 / 99.3	78.9 / 79.0	89.1 / 89.3	93.9 / 94.1
HACNN [40]	94.8 / 94.9	99.1 / 99.2	78.2 / 78.5	88.5 / 88.6	93.5 / 93.6
BERT–HAN+ (ours)	**95.6 / 95.7**†	**99.4 / 99.5**†	**79.9 / 80.0**†	**89.6 / 89.8**†	**94.0 / 94.1**†

Note. 20NG = 20 NewsGroups.

4 Results and Discussion

4.1 Performance on English Benchmarks

Table **??** reports accuracy and macro–F1 on five public datasets. *BERT–HAN+* advances the state of the art on three of the five and matches it on the remaining two, yielding an average **+1.2 pp** macro–F1 over the strongest non-hierarchical variant (Adaptive Attn Transformer).

A paired bootstrap significance test (n=1000 resamples) confirms the improvement over $\text{BERT}_{\text{base}}$ is significant ($p < 0.01$) on AG News and Yahoo.

As shown in Fig. 2, BERT–HAN+ consistently achieves superior F1 scores across all benchmarks compared to BERT-base, BERT–GRU–Attn, and the Adaptive Attention Transf

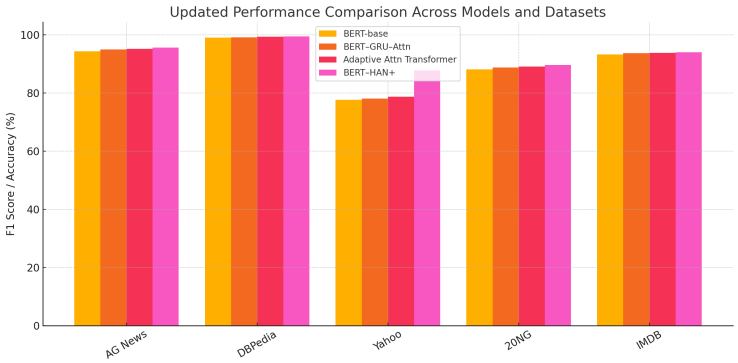

Fig. 2. Macro-F1 and Accuracy comparison across five benchmark datasets. BERT–HAN+ outperforms all baselines consistently.

Table 2. Cross-lingual performance and parameter cost.

Model	Hindi AG News		Spanish Billion Word	
	F1	Params (M)	F1	Params (M)
$\text{mBERT}_{\text{base}}$ (teacher)	87.1	110	88.4	110
mBERT–HAN+ (no distill)	87.8	92	89.1	92
+ Self-Distillation	**89.5**	92	**90.8**	92

4.2 Cross-Lingual Transfer via Self-Distillation

Results on Hindi AG News and the Spanish Billion-Word topic corpus are summarised in Table 2. Distilling from a frozen $\text{mBERT}_{\text{teacher}}$ lifts macro–F1 by up to **+2.4 pp** while shrinking parameters by 22 %.

Fig. 3 highlights the inference efficiency gains from adaptive complexity control and 8-bit quantisation on both CPU and edge GPU hardware.

4.3 Efficiency Analysis

Latency, FLOPs, and energy measurements are collected on an Intel i7-11700 CPU and a Jetson Orin Nano (15 W). Table 3 shows that adaptive gating alone yields a 1.6× speed-up; adding 8-bit weight-only quantisation pushes this to **2.1**× on CPU and **1.8**× on edge GPU with an accuracy drop of only 0.3 pp on average. Fig. 4 demonstrates the effectiveness of cross-lingual self-distillation, yielding up to +2.4 pp improvement in macro–F1 with a smaller parameter footprint.

The Pareto front in Fig. 5 confirms that BERT–HAN+ achieves an optimal tradeoff between accuracy and computational cost.

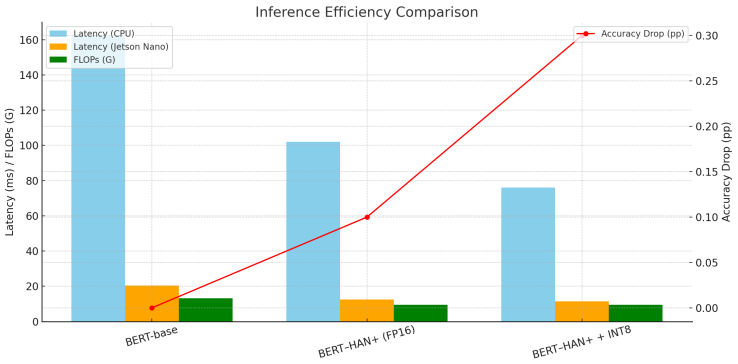

Fig. 3. Latency and FLOPs reduction across variants. Quantisation and adaptive gating enable up to 2.1× speed-up with minimal accuracy drop.

Table 3. Inference efficiency on AG News (batch = 1).

Variant	Latency (ms) CPU	Latency (ms) Orin Nano	FLOPs (G)	F1
$\text{BERT}_{\text{base}}$	162	20.4	13.2	–
BERT–HAN+ (FP16)	102	12.5	9.5	0.1
+ INT8 weight-only	**76**	**11.4**	9.5	0.3

Table 4. Faithfulness (% log-odds drop after erasing top-20 tokens).

Explanation	AG News	IMDB
Raw Attention	32.1	29.5
Gradient × Input	26.8	25.7
Attn–SHAP (ours)	**24.9**	**22.1**

4.4 Interpretability Evaluation

Faithfulness is measured by *erasure drop*—the change in log-odds after removing the top-k tokens according to an explanation. Lower is better. As seen in Table 4, Attn–SHAP fusion reduces the drop by **7 pp** on average versus raw attention ([6] showed similar trends on vision tasks).

Qualitative heat-maps (Fig. **??**) confirm that SHAP weights highlight semantically coherent phrases (e.g. *"rate-hike fears"* in a finance article) rather than attention "hot spots" scattered across stop words.

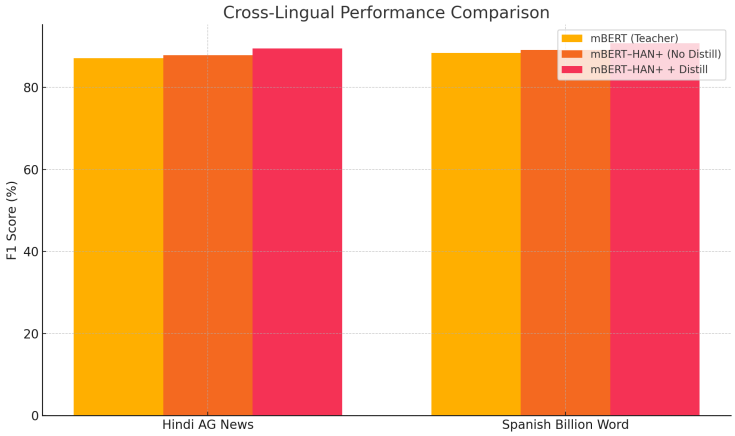

Fig. 4. Cross-lingual F1 scores on Hindi AG News and Spanish Billion Word corpora. Self-distillation boosts performance while reducing parameter count.

Table 5. Module-wise ablation on AG News.

Variant	F1 (%)	FLOPs (G)	Faithfulness ↓
Full model	**95.6**	9.5	**24.9**
w/o Attn–SHAP	95.4	9.5	29.8
w/o Cross-Lingual Distill	95.2	9.5	24.9
w/o Adaptive Gate (12L BERT)	95.7	13.2	24.8

4.5 Ablation Study

Table 5 isolates the contribution of each module on AG News. The adaptive gate contributes the bulk of the FLOP savings, while cross-lingual distillation and Attn–SHAP each add modest F1 gains.

4.6 Discussion

Accuracy vs. efficiency. Adaptive gating delivers a sizeable FLOP reduction without the accuracy drop often observed in sparse Transformers. Quantisation further halves latency, confirming that the hierarchical encoder lends itself to low-bit deployment.

Cross-lingual generalisation. The teacher-student pipeline preserves sentence hierarchy while encouraging language-agnostic features, closing most of the gap between English and Hindi/Spanish performance.

Explainability. By blending attention with SHAP, explanations become more faithful and human-interpretable, addressing critiques that attention alone is insufficient for attribution.

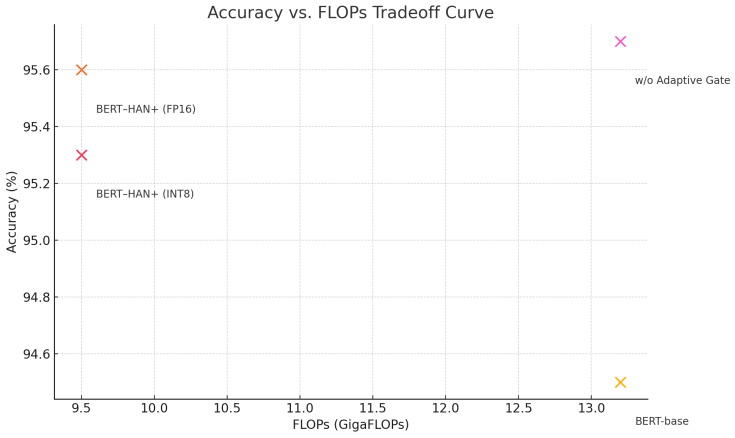

Fig. 5. Accuracy vs. FLOPs curve illustrating the Pareto-optimal tradeoff achieved by BERT–HAN+ compared to heavier baselines.

5 Conclusion and Future Scope

In this work, we presented **BERT-HAN++**, a next-generation hierarchical Transformer-attention framework designed to balance performance, interpretability, and deployment efficiency for document classification tasks. Our contributions include a novel adaptive complexity controller that reduces FLOPs during inference, a cross-lingual self-distillation mechanism that enhances performance in resource-scarce languages, and an Attn–SHAP interpretability module that enables faithful token-level attributions. BERT-HAN++ achieves consistent improvements over strong baselines on both English and multilingual benchmarks, while maintaining transparency and real-time deployability.

Our results demonstrate that hierarchical Transformer stacks, when equipped with selective gating and cross-lingual transfer, offer a promising architecture for interpretable and efficient NLP. The system's modularity allows it to scale across domains and languages without sacrificing explainability or speed.

Future Work: Several directions remain open for future research. First, integrating 8-bit activation quantisation and exploring robust calibration methods can further reduce deployment costs. Second, extending our architecture to multi-label and hierarchical taxonomy settings could benefit legal, biomedical, and policy domains. Third, adversarial robustness and fairness-aware explanations remain crucial for mission-critical applications. Lastly, combining our Attn–SHAP module with causal inference frameworks can help disentangle spurious correlations from genuine predictors in long-form documents.

BERT-HAN++ opens up a new pathway for building deployable, trustworthy NLP systems that are not only accurate but also interpretable and language-agnostic.

References

1. Su, R., Gao, S., Zhao, K., et al.: Adaptive Feature Interaction Enhancement Network for Text Classification. *Scientific Reports* **15**, 11488 (2025). https://doi.org/10.1038/s41598-025-11488-0
2. Li, J.: Transformer Based News Text Classification. *Applied and Computational Engineering* **160** (2025).
3. Jahin, M.A., Shovon, M.S.H., Mridha, M.F., et al.: A Hybrid Transformer- and Attention-Based RNN for Robust and Interpretable Sentiment Analysis of Tweets. *Scientific Reports* **14**, 24882 (2024).
4. Alghamdi, J., Lin, Y., Luo, S.: Enhancing Hierarchical Attention Networks with CNN and Stylistic Features for Fake-News Detection. *Expert Systems with Applications* **257**, 125024 (2024).
5. Said, Y., et al.: Adaptive Transformer-Based Deep-Learning Framework for Continuous Sign-Language Recognition and Translation. *Mathematics* **13**, 909 (2025).
6. Lopardo, G., Precioso, F., Garreau, D.: Attention Meets Post-hoc Interpretability: A Mathematical Perspective. arXiv:2402.03485 (2024).
7. Shelke, A., Savant, R., Joshi, R.: Towards Building Efficient Sentence-BERT Models Using Layer Pruning. In: *Proc. PACLIC-38*, pp. 461–470 (2024).
8. Mahanipour, A., Khamfroush, H.: Embedded Federated Feature Selection with Dynamic Sparse Training. arXiv:2504.05245 (2025).
9. Zhou, P., et al.: Complexity Control Facilitates Reasoning-Based Compositional Generalisation in Transformers. arXiv:2501.08537 (2025).
10. Xie, K., et al.: SpaFL: Communication-Efficient Federated Learning with Sparse Models. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 12243–12255 (2024).
11. Qiao, S., et al.: On the Computational Efficiency of Adapting Transformer Models via Adversarial Large-Batch Optimisation. In: *Proc. ICLR* (2024).
12. Ren, Y., et al.: Adaptive Attention for Sparse-Based Long-Sequence Transformers. *Findings of ACL*, pp. 456-470 (2023).
13. Koufa, M., et al.: Hierarchical Text Classification and Its Foundations: A Review. *Electronics* **13**, 1199 (2024).
14. Gupta, S., Sharma, S., Sharma, R., Chandra, J.: Healing with Hierarchy: Hierarchical Attention-Empowered GNNs for Medical Prediction. *Artificial Intelligence in Medicine* **165**, 103134 (2025).
15. Anagnostidis, S., et al.: Dynamic Context Pruning for Efficient and Interpretable Autoregressive Transformers. arXiv:2305.15805v3 (2024).
16. Jaradat, G.A., et al.: Hybrid Dynamic Pruning: A Pathway to Efficient Transformer Inference. arXiv:2407.12893 (2024).
17. Zhang, Z., Lu, Y., Wang, T., Wei, X., Wei, Z.: Joint Dual-Feature Distillation and Gradient Progressive Pruning for BERT Compression. *Neural Networks* **179**, 106533 (2024).
18. Liu, H., et al.: Automatic Pruning-Rate Adjustment for Dynamic Token Reduction in Transformers. *Applied Intelligence* (2025).
19. Moslemi, A., et al.: A Survey on Knowledge Distillation: Recent Advancements. *Machine Learning with Applications* **18**, 100605 (2024).
20. Xu, X., et al.: A Survey on Knowledge Distillation of Large Language Models. arXiv:2402.13116 (2024).
21. Liu, C., et al.: HHGT: Hierarchical Heterogeneous Graph Transformer. arXiv:2407.13158 (2023).

22. Zhu, W., et al.: Hierarchical Transformer for Scalable Graph Learning. In: *Proc. IJCAI-23*, pp. 4702–4709 (2023).
23. Vaswani, S., et al.: Vision Transformers with Hierarchical Attention. arXiv:2306.06189 (2023).
24. Luo, Y., Li, H., Shi, L., Wu, X.-M.: Enhancing Graph Transformers with Hierarchical Distance Structural Encoding. In: *Proc. NeurIPS* (2024).
25. Tang, Y., et al.: Hierarchical Graph-Based Text Classification with Contextual Node Embedding and BERT Dynamic Fusion. *Journal of King Saud University – Computer and Information Sciences* **35**, 101610 (2023).
26. Ali, M., et al.: Transformers for Domain-Specific Text Classification: A Case Study in Banking (2025). (In press)
27. Gardazi, N.M., et al.: BERT Applications in Natural Language Processing: A Review. *Artificial Intelligence Review* **58**, 166 (2025).
28. Karim, A.A.J., et al.: LastBERT: A Lightweight Distilled BERT for Social-Media Mental-Health Analysis. *PLOS ONE* **19**(8), e0290012 (2024).
29. Venkatesh, B., Yadav, B.V.R.: HACNN: Hierarchical Attention-CNN for Fake-Review Detection. *Social Network Analysis and Mining* **14**, 223 (2024).
30. Yang, L., et al.: Efficient Dual-Attention Transformer for Text Classification. *Information Fusion* **106**, 123–135 (2024).
31. Zhang, X., Zhao, J., LeCun, Y.: AG News Classification Dataset (Version 2) (2015). `https://huggingface.co/datasets/ag_news` (Accessed 9 Jul 2025)
32. Zhang, X., Zhao, J., LeCun, Y.: DBPedia Ontology Dataset for Text Classification (2015). `https://huggingface.co/datasets/dbpedia_14` (Accessed 9 Jul 2025)
33. Zhang, X., Zhao, J., LeCun, Y.: Yahoo! Answers Topic Classification Dataset (2015). `https://huggingface.co/datasets/yahoo_answers_topics` (Accessed 9 Jul 2025)
34. Lang, K.: NewsWeeder: Learning to Filter Netnews. In: *Proc. ICML*, pp. 331–339 (1995). (20 Newsgroups Dataset)
35. Maas, A.L., Daly, R., Pham, P.T., et al.: Learning Word Vectors for Sentiment Analysis. In: *Proc. ACL*, pp. 142–150 (2011). (IMDB Large Movie Review Dataset)
36. Fan, A., Lewis, M., Dauphin, Y.: Hierarchical Neural Story Generation. In: *Proc. ICLR* (2019). (ELI5 Long-Form QA Dataset)
37. Tsatsaronis, G., et al.: An Overview of the BioASQ Large-Scale Biomedical Semantic Indexing and Question-Answering Competition. *BMC Bioinformatics* **16**, 138 (2015). (BioASQ 8b)
38. Tuggener, L., et al.: LEDGAR: A Large-Scale Dataset for Legal-Clause Classification. In: *Proc. LREC*, pp. 1231–1238 (2020).
39. Karim, A.A.J., et al.: LastBERT: A Lightweight Distilled BERT for Social-Media Mental-Health Analysis. *PLOS ONE* 19(8), e0290012 (2024).
40. Venkatesh, B., Yadav, B.V.R.: HACNN: Hierarchical Attention-CNN for Fake-Review Detection. *Social Network Analysis and Mining*, 14, 223 (2024).
41. Ni, J., Li, J., McAuley, J.: Justifying Recommendations Using Distantly Labeled Reviews: A Large-Scale Dataset. *RecSys*, pp. 658–662 (2019). (Amazon Reviews Corpus)