

Enhancing Burmese News Classification with Kolmogorov-Arnold Network Head Fine-tuning

Thura Aung^{1,2}, Eaint Kay Khaing Kyaw^{1,2}, Ye Kyaw Thu^{1,3*}, Thazin Myint Oo¹, Thepchai Supnithi^{3*}

¹*Language Understanding Laboratory, Myanmar*

²*Department of Computer Engineering, KMITL, Bangkok, Thailand*

³*Language and Semantic Technology Research Team, NECTEC, Bangkok, Thailand*

Emails: 66011606@kmitl.ac.th, 66011533@kmitl.ac.th,

{yekyaw.thu, thepchai.supnithi}@nectec.or.th, queenofthazin@gmail.com

Abstract—In low-resource languages like Burmese, classification tasks often fine-tune only the final classification layer, keeping pre-trained encoder weights frozen. While Multi-Layer Perceptrons (MLPs) are commonly used, their fixed non-linearity can limit expressiveness and increase computational cost. This work explores Kolmogorov–Arnold Networks (KANs) as alternative classification heads, evaluating Fourier-based FourierKAN, Spline-based EfficientKAN, and Grid-based FasterKAN—across diverse embeddings including TF-IDF, fastText, and multilingual transformers (mBERT, Distil-mBERT). Experimental results show that KAN-based heads are competitive with or superior to MLPs. EfficientKAN with fastText achieved the highest F1-score (0.928), while FasterKAN offered the best trade-off between speed and accuracy. On transformer embeddings, EfficientKAN matched or slightly outperformed MLPs with mBERT (0.917 F1). These findings highlight KANs as expressive, efficient alternatives to MLPs for low-resource language classification.

Index Terms—Kolmogorov-Arnold Network, Head finetuning, Transformers, News Classification, FourierKAN, FasterKAN, EfficientKAN

I. INTRODUCTION

News sentence classification involves assigning predefined categories to news text by analyzing its content and identifying meaningful patterns. Automating this task enables faster and more accurate handling of large-scale news data. In this work, we investigate head fine-tuning with different classification heads. Using the standard multilayer perceptron (MLP) head as a baseline, we evaluate three variants of the Kolmogorov–Arnold Network (KAN) [1]—FourierKAN [2], EfficientKAN [3] and FasterKAN [4]. These are tested across both static embeddings, including TF-IDF, random, and fastText [5], and contextual embeddings derived from transformer backbones - multilingual BERT (mBERT) [6] and Distil-mBERT [7].

We contributed Burmese News Classification Dataset and presented the performance comparison between MLP and KAN classification heads. Our results show that EfficientKAN combined with fastText embeddings

achieves the best overall performance (F1 of 0.928), while FasterKAN provides the fastest training with competitive accuracy. We further analyze trade-offs between accuracy, parameter size, and training efficiency, demonstrating that KAN-based heads are lightweight yet expressive alternatives to conventional MLPs in low-resource settings.

II. RELATED WORK

Previous studies on Burmese news classification have applied traditional machine learning models [8], [9] as well as deep learning approaches, including CNN [10], BiLSTM [11], and hybrid models [12], [13], using both syllable- and word-level tokenization [14].

A widely used approach for adapting pre-trained models to downstream tasks is classification head fine-tuning, often referred to as linear probing. In this setup, only the final classification layer is updated while the underlying backbone remains fixed. This strategy is particularly effective in low-resource scenarios [15] and has been shown to maintain stronger robustness under out-of-distribution shifts [16] compared with full-model fine-tuning.

In natural language processing tasks such as text classification, it is common to place a Multi-Layer Perceptron (MLP) head [17] on top of transformer-based backbones [18]. Despite their popularity, MLP heads may fail to fully exploit the contextual richness of pre-trained embeddings and often introduce significant computational overhead. Recent studies [19] have proposed FourierKAN, a KAN variant, which leverage Fourier-based transformations and have demonstrated superior performance over conventional MLP heads in fine-tuning tasks.

III. DATASET PREPARATION

We collected Burmese news paragraphs across six categories from various news websites and manually labeled them, as summarized in Table II. The news articles were gathered from Burmese news websites such as VOA Burmese¹, BBC Burmese², and Radio Free Asia (RFA)³, and annotated by native speakers between April and June

*Corresponding author.

Preprint; Accepted to iSAI-NLP 2025.

¹<https://burmese.voanews.com/>

²<https://www.bbc.com/burmese>

³<https://www.rfa.org/burmese/>

TABLE I: Example Sentences from Each News Type from News Classification Dataset.

Class	Example Sentences
Sports	အကြံ့မိတ်လွှဲပွဲ တွင် အဆင့် ရှု ဖို့ သည့် နှစ် သင်း က မိတ်လွှဲပွဲ ထပ်မံ ယဉ်ယူ၏ ရ မည် ဖြစ် ပြီး ရှုံးနိုင် သည့် နှစ် သင်း က တတိယ နေ ရာ လွှဲ ဆက်လက် ကဲး၊ ရ မည့် ဖြစ် သည်။ The two teams that win the semi-finals will compete in the final, while the two losing teams will play in the third-place match.
Politics	အရေးအကြီးဆုံး က သူ ကို ခြေား မှု၊ အလွှာသုံးစား မှု နဲ့ အမြတ်ထုတ် မှု စွဲပွဲ မှု တွေ မ ရှိ ဖူး သေး ပါ။ မစွဲတာ အူမဲ့ရော် ဟာ ဟောလော် ပါတီ ကနေ ရွှေးကောက်ပွဲ ဝင် ရင်း နှင့် ရေး လောက ထဲ ကို ပျော် မှု ဝင်ရောက လာ ခဲ့ တယ်။ အဲဒီ နောက် အေးရှု အရော့၍ ဖြစ် လာ ခဲ့ ပါ တယ်။ Most importantly, he has never been accused of corruption, misuse of funds, or exploitation. Mr. Umerov entered politics in 2019, running as a candidate for the Holos party. After that, he became a government official.
Technology	နှင့်တကာ အာကာသ စခန်း ဟာ လာ မယ် ၂၀၃၀ အထိ အလုပ် လုပ် နေ ဦး မှာ ဖြစ် ပြီး ၂၀၃၁ မှာ ပစ္စိတ် သမုဒ္ဒရာ ထဲ ကို ပျက်ကျ လာ လိမ့် မယ် လို့ နာဆာ က ပြော ပါ တယ်။ NASA says the International Space Station will continue to operate until 2030 and will de-orbit into the Pacific Ocean in 2031.
Business	၂၀၂၃ ၂၀၂၄ ဘဏ္ဍာ နှစ် ပါး လ အတွင်း ပြည်ပ သို့ ရေ ထွက် ပစ္စိုး တင် ပို မှ မှ ဒေါ်လာ ၂၄၁ သန်း ကျော် ရ ရှိ ပြီး ပင်လယ် ရေ ကြောင်း ကန်သို့ မှ မှ ဒေါ်လာ ၁၃၄ သန်း ကျော် တင် ပို ထား ရောင်းရန်းရောင်းဝယ် ရေး စိုးပြီးဌာန မှ သို့ ရ သည်။ The Ministry of Commerce announced that over \$241 million was earned from exporting fishery products abroad during the first five months of the 2023–2024 fiscal year, with more than \$134 million of that coming from maritime trade.
Entertainment	အဖွဲ့ ဝင် လီော် အင်စတာဂုဏ် ထွင် ပိုစ် တစ် ခု တင် လိုက် သည် နှင့် အမောက်နှင့် ဒေါ်လာ ၅၇၅,၀၀၀ ရ နေ ပြီး သား ဟု ဆိုရယ်မီဒီယာ မားကတ်တင်း ပောက်ဖောင်း က ထုတ်ပြု သည့် အင်စတာဂုဏ် ပြင် ချမ်းသာ နေ သူ များ စာရင်း တွင် ဆို ထား သည်။ According to a list of "Instagram rich" compiled by a social media marketing platform, group member Lisa earns \$575,000 for each post on Instagram.
Environment	ဒီ မျောက် တွေ ကို အမဲ လိုက် သတ်ဖြတ် နေ တာ နဲ့ ပက်သက် ပြီး ဥပဒေ အရ အရေးယူ တာ တွေ၊ မျောက် လေး တွေ ကို ထိန်းသိမ်း တာ တွေ ကို တော့ မ ကြား မိ သေး ဘူး လို ကိုင်းဝိုင်း က ဆက် ပြော ပါ တယ်။ Ko Win Paing Oo added that he hasn't heard of any legal action being taken against those who are hunting and killing the monkeys or any efforts to conserve them.

TABLE II: Label Counts and Percentages

Class	Count	Percentage
Sports	1,232	16.84%
Politics	1,228	16.79%
Technology	1,224	16.73%
Business	1,221	16.69%
Entertainment	1,205	16.47%
Environment	1,205	16.47%

2024. The dataset was split into training 80% (5.84k sentences) and testing 20% (1.47k sentences) of the original dataset.

Prior to experimentation, the text underwent cleaning and normalization based on syllable rules to ensure correct Unicode ordering. Proper ordering is crucial for accurate word and sentence segmentation [20], [21] and for consistent model training. After preprocessing, we performed tokenization using mBert tokenizer for transformer-based approaches. For static embedding-based models, we used the myWord⁴ [22] tool, which utilizes unigram and bigram dictionaries derived from 0.5 million sentences and 12 million words. Figure I shows examples of each class in our news classification dataset.

IV. METHODOLOGY

A. Embeddings

In this work, we focus on evaluating the effectiveness of different embedding strategies combined with different classification heads. Neural embeddings map discrete tokens into continuous vectors that capture semantic and

⁴<https://github.com/ye-kyaw-thu/myWord>

syntactic relationships. We compare static embeddings and contextual embeddings while keeping the backbone frozen and fine-tuning only the classification head.

For static embeddings, we experiment with TF-IDF, random embeddings, and fastText⁵ [5] embeddings. These embeddings provide fixed representations for each token, which are then fed into a classification head. For contextual embeddings, we use pre-trained models such as multilingual BERT (mBERT⁶) [6] and multilingual distilled Bert (Distil-mBERT⁷) [7].

These embeddings dynamically encode each token based on its context in the input sequence and are also passed through a classification head without further updating the backbone. As illustrated in Figure 1, the input text is first processed by a pretrained Transformer encoder to obtain contextual embeddings for each token. These embeddings are then passed to the KAN layer, which is trained to serve as the classification head while keeping the encoder backbone fixed. The trained KAN layer maps the contextual representations to their corresponding predicted classes.

B. Models

1) *Baseline MLP*: We use a standard Multi-Layer Perceptron (MLP) as a baseline classifier. For a 1-layer MLP:

$$\hat{y} = \text{softmax}(W_0H + b_0) \quad (1)$$

and for a 2-layer MLP:

$$h_0 = \sigma(W_0H + b_0), \quad (2)$$

$$\hat{y} = \text{softmax}(W_1h_0 + b_1), \quad (3)$$

⁵facebook/fasttext-my-vectors

⁶google-bert/bert-base-multilingual-cased

⁷distilbert/distilbert-base-multilingual-cased

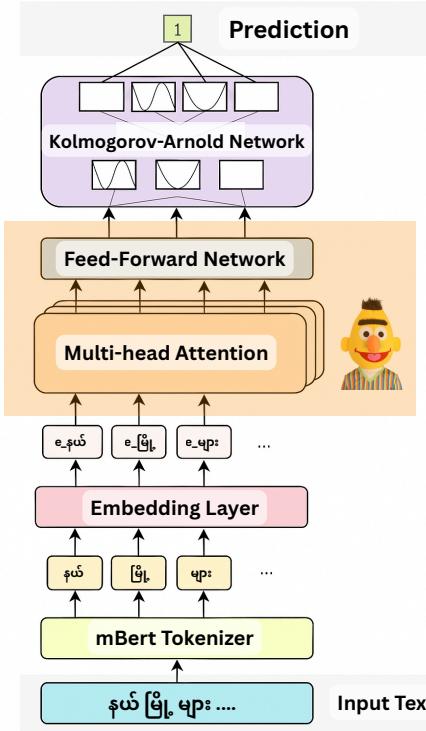


Fig. 1: Integration of KAN as a classification head on top of a Transformer encoder for context-aware prediction

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid activation, H is the input embedding, and W_i, b_i are trainable parameters.

2) *Kolmogorov-Arnold Representation Theorem (KART)*: KART [23], [24] states that any multivariate continuous function can be decomposed into sums of univariate functions:

$$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^n \phi_{q,p}(x_p) \right), \quad (4)$$

where $\phi_{q,p}$ are simple single-variable functions applied to each input x_p , and Φ_q are outer functions combining these sums. This decomposition motivates KANs, which use learnable $\phi_{q,p}$ and Φ_q for adaptive modeling.

3) *Kolmogorov-Arnold Network (KAN)*: KAN layers extend KART by parameterizing $\phi_{q,p}$ and Φ_q as learnable splines. For a single KAN layer with n_{in} input and n_{out} output dimensions, the layer output is:

$$\mathbf{x}_j^{(l+1)} = \sum_{i=1}^{n_l} \phi_{j,i}^{(l)}(\mathbf{x}_i^{(l)}), \quad j = 1, \dots, n_{l+1}, \quad (5)$$

or in matrix form:

$$\mathbf{x}^{(l+1)} = \Phi^{(l)} \mathbf{x}^{(l)} \quad (6)$$

where $\Phi^{(l)}$ is the function matrix of learnable splines for the l -th layer. The approximation error is bounded by:

$$\|f - (\Phi_G^{(L-1)} \circ \dots \circ \Phi_G^{(0)})\mathbf{x}\|_{C^m} \leq CG^{-k-1+m}, \quad (7)$$

with grid size G , spline order k , and constant C .

FourierKAN FourierKAN [19] replaces spline-based activations with Fourier series expansions to represent the univariate functions $\phi_{q,p}(x_p)$ in the Kolmogorov–Arnold decomposition. Each function is approximated as a weighted sum of sine and cosine terms:

$$\phi_{q,p}(x_p) = \sum_{k=1}^G a_{q,p,k} \cos(kx_p) + b_{q,p,k} \sin(kx_p), \quad (8)$$

where G is the grid size (number of Fourier modes), and $a_{q,p,k}, b_{q,p,k}$ are learnable Fourier coefficients. This formulation captures smooth periodic and non-linear patterns efficiently while maintaining a compact parameterization. The FourierKAN layer thus computes:

$$\mathbf{x}_j^{(l+1)} = \sum_{i=1}^{n_l} [A_{j,i}^{(l)} \cos(kx_i^{(l)}) + B_{j,i}^{(l)} \sin(kx_i^{(l)})] + \beta_j, \quad (9)$$

where β_j denotes an optional bias term. Compared to spline-based KANs, FourierKAN offers smoother gradient flow, easier initialization, and efficient modeling of oscillatory or compositional data.

Efficient KAN In spline-based KAN, sparsity, an important property for interpretability, was originally enforced via L1 regularization on the expanded input tensors [1]. In EfficientKAN, this is replaced with L1 regularization on the weights, maintaining computational efficiency while encouraging sparsity. The learnable scaling for each activation function can be optionally enabled or disabled, providing a trade-off between accuracy and efficiency.

FasterKAN The original KAN expands the input tensor to shape (batch_size, out_features, in_features) in order to apply learnable activations, which incurs high memory usage. EfficientKAN [3] addresses this by first applying the activation functions to the input and then performing a linear combination of the results. This reformulation significantly reduces memory consumption and allows forward and backward computations to be implemented as standard matrix multiplications. For the activation basis, the FasterKAN authors [4] investigate Reflectional Switch Activation Functions (RSWAF):

$$b_i(u) = 1 - \tanh^2 \left(\frac{u - u_i}{h} \right), \quad (10)$$

which approximate spline functions, are computationally inexpensive, and can be evaluated on uniform grids. In addition, we explore alternative basis functions, different exponent and h values, and the exclusion of SiLU in our experiments.

Finally, careful initialization of the weight parameters using Kaiming uniform initialization has been found to improve convergence and overall performance. Combining these strategies results in the FasterKAN variant, which integrates both computational efficiency and competitive predictive performance.

TABLE III: Efficiency comparison of models with different embeddings

Model	Params	Train (s)	Fwd (ms)	Bwd (ms)
Tf-IDF + MLP	0.13M	10.7	0.26	0.64
Tf-IDF + FourierKAN	2.99M	28.1	0.56	1.23
Tf-IDF + EfficientKAN	1.29M	21.4	4.49	4.18
Tf-IDF + FasterKAN	1.03M	7.6	0.58	1.00
Random Embedding + MLP	2.00M	13.5	0.44	0.97
Random Embedding + FourierKAN	0.27M	8.0	0.56	1.23
Random Embedding + EfficientKAN	2.35M	24.5	1.76	2.82
Random Embedding + FasterKAN	2.27M	14.2	0.78	1.41
fastText + MLP	2.00M	18.1	0.46	0.98
fastText + FourierKAN	0.63M	8.3	1.08	1.90
fastText + EfficientKAN	2.35M	24.6	1.82	2.90
fastText + FasterKAN	2.27M	15.6	0.75	1.31
Distil-mBERT + MLP	135M	669.5	100.23	211.76
Distil-mBERT + FourierKAN	137M	795.2	52.09	113.49
Distil-mBERT + EfficientKAN	137M	671.0	100.13	212.90
Distil-mBERT + FasterKAN	136M	669.4	99.92	211.73
mBERT + MLP	178M	1284.8	203.03	418.72
mBERT + FourierKAN	180M	1481.2	115.33	215.38
mBERT + EfficientKAN	180M	1291.2	203.57	420.01
mBERT + FasterKAN	179M	1289.8	202.75	418.26

V. EXPERIMENTAL SETUP

We conducted experiments on Google Colab using an NVIDIA Tesla T4 GPU with 16 GB VRAM. The models were implemented in PyTorch, with supporting libraries including scikit-learn for evaluation metrics and Hugging Face Transformers for pre-trained multilingual backbones.

For models based on static embeddings, we employed TF-IDF, random, and fastText representations. For models with transformer-backbones, we evaluated three multilingual architectures: mBERT, and Distil-mBERT. On top of each backbone, we integrated different classification heads to compare their effectiveness.

- **MLP**: Feed-forward network with ReLU activations.
- **FourierKAN**: Fourier-based KAN variant with grid size of 8, using sine and cosine basis functions.
- **EfficientKAN**: spline-based variant with grid size of 8 and cubic spline order.
- **FasterKAN**: grid-based KAN variant with learnable grid and inverse denominator formulation.

The training used the AdamW optimizer with a two-tier learning rate: 2e-5 for the backbone parameters and 2e-4 for the classification head, along with a Cosine Annealing scheduler. The cross-entropy loss function was used. Training was performed for 15 epochs with batch size 32 for static embeddings approaches and 5 epochs with batch size 8 for transformer-backbone models, and gradient clipping (max-norm = 1.0) was applied to stabilize training.

Model accuracy was assessed on the held-out test set, with weighted F1-score reported as the evaluation metric. To assess computational efficiency, we measured the num-

ber of parameters (total and trainable), training time, and average forward/backward pass latency (in milliseconds). Early stopping with patience of three validation checks and dropout rate of 0.3 was applied to avoid overfitting.

TABLE IV: F1-score comparison of classification heads across embeddings

Model	F1-score
Tf-IDF + MLP	0.783
Tf-IDF + FourierKAN	0.538
Tf-IDF + EfficientKAN	0.791
Tf-IDF + FasterKAN	0.560
Random Embedding + MLP	0.915
Random Embedding + FourierKAN	0.699
Random Embedding + EfficientKAN	0.917
Random Embedding + FasterKAN	0.911
fastText + MLP	0.908
fastText + FourierKAN	0.829
fastText + EfficientKAN	0.928
fastText + FasterKAN	0.927
Distil-mBERT + MLP	0.859
Distil-mBERT + FourierKAN	0.788
Distil-mBERT + EfficientKAN	0.864
Distil-mBERT + FasterKAN	0.873
mBERT + MLP	0.917
mBERT + FourierKAN	0.877
mBERT + EfficientKAN	0.917
mBERT + FasterKAN	0.913

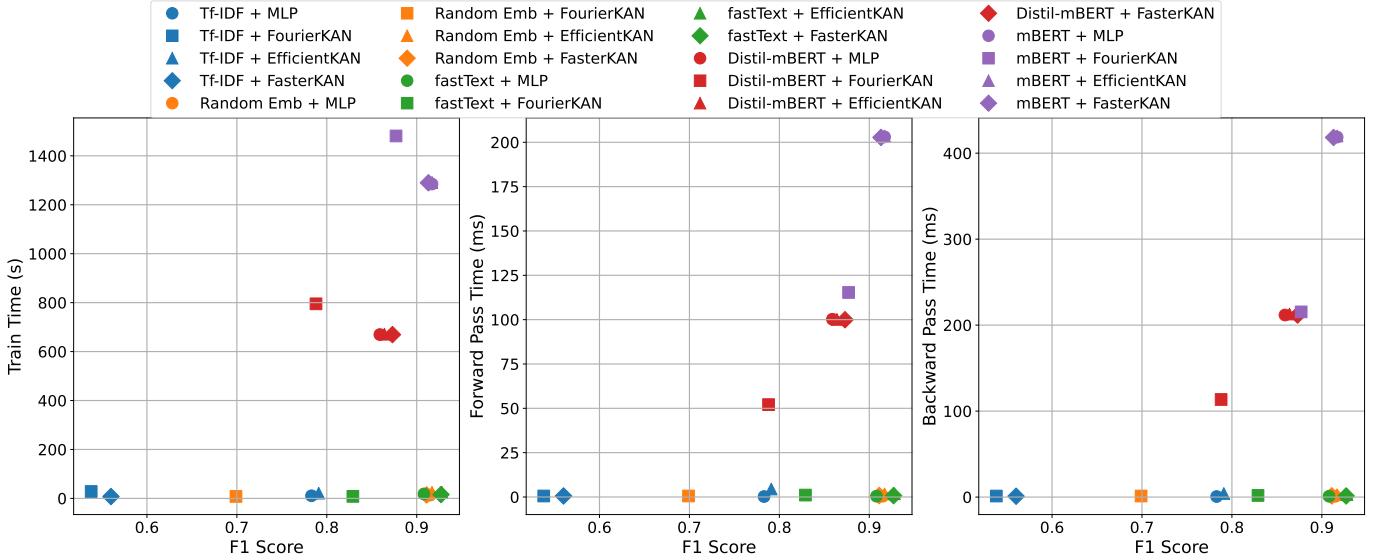


Fig. 2: Comparison of model performance versus computational cost.

VI. RESULT AND DISCUSSION

Figure 2 illustrates the trade-off between model performance and computational cost across various model and embedding configurations. Each subplot plots the F1-score on the x-axis against a different computational metric on the y-axis: training time, forward pass time, and backward pass time, respectively (with lower values indicating better efficiency). The color coding distinguishes embedding types — blue for TF-IDF baseline, orange for random learned embeddings, green for fastText embeddings, and red for multilingual Transformer embeddings — while marker shapes differentiate the classification heads: circles for MLP, squares for FasterKAN, triangles for EfficientKAN, and diamonds for FourierKAN.

A. Efficiency Comparison

Training Time Table III shows that MLP serves as the baseline and is generally the fastest to train across embeddings. Interestingly, FasterKAN often trains quicker than MLP despite having more parameters, indicating an efficient architecture. In contrast, EfficientKAN consistently requires significantly longer training times regardless of the embedding type. Transformer-based models, including Distil-mBERT and mBERT, are the slowest, with training durations ranging from several hundred seconds to over one thousand seconds.

Inference and Backward Propagation Efficiency Across all models, MLP consistently demonstrates the lowest latency during both forward inference (0.26–0.46 ms) and backward propagation (0.64–0.98 ms), making it the most computationally efficient choice. FourierKAN models generally have slightly higher times than MLP but remain efficient, with forward passes around 0.56–1.08 ms and backward passes between 1.23–1.90 ms depending

on the embedding. FasterKAN introduces moderate overheads, with forward times of 0.58–0.78 ms and backward times of 1.0–1.41 ms, while EfficientKAN is the slowest among non-transformer models, ranging from 1.76 to 4.49 ms forward and up to 4.18 ms backward. Transformer-based models incur substantially higher costs, with Distil-mBERT forward and backward passes around 52–100 ms and 113–212 ms respectively, and mBERT exceeding 115 ms forward and 215–418 ms backward, reflecting their greater computational demands during inference.

B. Performance Comparison

Effect of Classification Head Table IV shows EfficientKAN generally achieves the highest F1-scores for classical embeddings, followed by MLP and FasterKAN, with FourierKAN trailing behind. For TF-IDF, EfficientKAN scores 0.791 versus FourierKAN’s 0.538. Similar trends hold for Random and fastText embeddings. For transformers, FasterKAN or MLP tend to perform best, with FourierKAN slightly lower but competitive.

Effect of Embedding Type Embedding choice strongly affects performance. Learned embeddings (Random, fastText, Fourier) consistently outperform TF-IDF across all heads. fastText usually yields the highest F1, followed by Random and Fourier. Transformer embeddings provide additional gains, but improvements are smaller compared to switching from TF-IDF to learned embeddings.

Overall, FourierKAN generally underperforms compared to spline-based EfficientKAN and grid-based FasterKAN in terms of F1-score, due to architectural limitations. Its use of fixed Fourier basis functions reduces its ability to model complex, localized patterns, whereas spline-based KANs learn adaptive piecewise functions that better approximate nonlinear decision boundaries. While FourierKAN often has more trainable parameters and

longer training times than FasterKAN, it is typically faster in inference due to the simplicity of its basis functions. In contrast, FasterKAN uses a grid-based architecture with RSWAF to efficiently approximate B-splines, offering a strong balance between accuracy and efficiency. Overall, despite its speed in inference, FourierKAN’s lower accuracy and higher training cost make it less favorable for most classification tasks.

VII. CONCLUSION AND FUTURE WORK

This work explored Kolmogorov-Arnold Networks (EfficientKAN, FasterKAN, and FourierKAN) as classification heads for Burmese news sentence classification using static and contextual embeddings.

Our results show that embedding type has the strongest impact on performance. Learned embeddings like random and fastText outperform TF-IDF, while mBERT and Distil-mBERT achieve the highest overall accuracy. Among classification heads, EfficientKAN consistently performs best with static embeddings, especially fastText (0.928 F1), while FasterKAN offers a strong balance of speed and accuracy. FourierKAN, though efficient, underperforms due to its less adaptive basis and larger parameter counts. For transformer embeddings, differences between heads are smaller, with MLP, EfficientKAN, and FasterKAN performing similarly. MLP remains fastest overall, while EfficientKAN is the most computationally expensive among non-transformer heads.

KAN-based heads show promise, particularly with static embeddings. Future work will extend these models to other tasks in low-resource settings, explore new basis functions, and release code and data for reproducibility.

REFERENCES

- [1] Z. Liu *et al.*, “Kan: Kolmogorov-arnold networks,” *ArXiv*, 2024, accessed 11 Feb. 2025. [Online]. Available: <https://arxiv.org/abs/2404.19756>
- [2] GistNoesis, “Fourierkan,” <https://github.com/GistNoesis/FourierKAN>, 2025, GitHub repository, MIT License.
- [3] Blealtan, “efficient-kan: An efficient pure-pytorch implementation of kolmogorov-arnold network (kan),” <https://github.com/Blealtan/efficient-kan>, 2024, accessed: 2025-08-18.
- [4] A. Delis, “Fasterkan,” <https://github.com/AthanasisDelis/faster-kan/>, 2024.
- [5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.
- [6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: <http://arxiv.org/abs/1810.04805>
- [7] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” *ArXiv*, vol. abs/1910.01108, 2019.
- [8] A. H. Khine, K. T. Nwet, and K. M. Soe, “Automatic myanmar news classification,” in *Proceedings of the 15th International Conference on Computer Applications*, Feb. 2017, pp. 401–408.
- [9] K. T. Nwet, “Machine learning algorithms for myanmar news classification,” *International Journal on Natural Language Computing (IJNLC)*, vol. 8, no. 4, pp. 17–24, Aug. 2019.
- [10] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751.
- [11] J. Zhang, Y. Li, J. Tian, and T. Li, “Lstm-cnn hybrid model for text classification,” in *Proceedings of the 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Oct. 2018, pp. 1675–1680.
- [12] K. A. Kwaik, M. Saad, S. Chatzikyriakidis, and S. Dobnik, “Lstm-cnn deep learning model for sentiment analysis of dialectal arabic,” in *Proceedings of the International Conference on Arabic Language Processing*. Springer, Cham, Oct. 2019, pp. 108–121.
- [13] M. S. Phyu and K. T. Nwet, “A study on a joint deep learning model for myanmar text classification,” in *Proceedings of 2020 IEEE Conference on Computer Applications (ICCA)*, Yangon, Myanmar, 2020, pp. 1–4.
- [14] M. Phyu and K. T. Nwet, “Articles classification in myanmar language,” in *Proceedings of the 2019 International Conference on Advanced Information Technologies (ICAIT)*. Yangon, Myanmar: IEEE, Nov. 2019, pp. 188–193.
- [15] M. Gao, Q. Wang, Z. Lin, P. Zhu, Q. Hu, and J. Zhou, “Tuning pre-trained model via moment probing,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11803–11813.
- [16] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, “Fine-tuning can distort pre-trained features and underperform out-of-distribution,” *arXiv preprint arXiv:2202.10054*, 2022.
- [17] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.
- [18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, . Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [19] A. A. Imran and M. F. Ishmam, “FourierKAN outperforms MLP on text classification head fine-tuning,” in *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*, 2024. [Online]. Available: <https://openreview.net/forum?id=xqNDuVxThU>
- [20] T. Aung, Y. Kyaw Thu, and Z. Hlaing, “mysentence: Sentence segmentation for myanmar language using neural machine translation approach,” *Journal of Intelligent Informatics and Smart Technology*, vol. 9, no. October, p. e001, Nov. 2023.
- [21] Y. K. Thu, T. Aung, and T. Supnithi, “Neural sequence labeling based sentence segmentation for myanmar language,” in *Proceedings of the 12th Conference on Information Technology and Its Applications (CITA 2023)*, ser. Lecture Notes in Networks and Systems, vol. 734. Danang, Vietnam: Springer, Cham, Jul 28-29 2023, pp. 285–296.
- [22] Y. K. Thu, “myWord: Syllable, Word and Phrase Segmenter for Burmese,” <https://github.com/ye-kyaw-thu/myWord>, Sep. 2021, [Online; accessed 13-Jun-2024].
- [23] J. Schmidt-Hieber, “The kolmogorov-arnold representation theorem revisited,” *Neural Networks*, 2021.
- [24] S. Selitskiy, “Kolmogorov’s gate non-linearity as a step toward much smaller artificial neural networks,” in *Proceedings of the 24th International Conference on Enterprise Information Systems (ICEIS 2022)*. University of Bedfordshire, School of Computer Science and Technology, 2022.