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

Thura Aung, Eaint Kay Khaing Kyaw,

Ye Kyaw Thu, Thazin Myint Oo, Thepchai Supnithi

Language Understanding λ(μ) Laboratory

King Mongkut's Institute of Technology Ladkrabang

NECTEC
a member of NSTDA

Outline

- Introduction
- Related Work
- Dataset Preparation
- Methodology
- Experimental Setup
- Result and Discussion
- Conclusion and Future Work

Introduction

- News sentence classification assigns predefined categories to news text.
- Automating classification improves speed and accuracy for large-scale news streams.

  This work investigates head fine-tuning with different classification heads.
- Baseline: standard MLP classification head.
- Compared with three KAN variants: FourierKAN, EfficientKAN, FasterKAN.
- Experiments conducted on both:
  - Static embeddings: TF-IDF, random, fastText
  - Contextual embeddings: mBERT, Distil-mBERT

Introduction (Cont'd)

- Introduced a Burmese News Classification Dataset.
- Compared performance of MLP vs. KAN-based classification heads.
- EfficientKAN + fastText achieves best overall F1 score (0.928).
- FasterKAN provides fastest training with competitive accuracy.
- KAN heads show strong trade-offs between accuracy, parameter size, and training cost.
- Results indicate KAN-based heads are lightweight yet expressive alternatives in low-resource settings.

Related Work

- Prior work used traditional ML models, CNN, BiLSTM, and hybrid neural models for classification.
- Experiments conducted with both syllable-level and word-level tokenization.

Classification Heads and Fine-Tuning

- Linear probing updates only the classification head while keeping the backbone frozen.
- Effective for low-resource settings and robust under distribution shifts.
- Standard practice: attach an MLP head on transformer-based backbones.
- MLP heads remain widely used but may underutilize rich contextual embeddings.
- Recent work proposes KAN-based heads (using FourierKAN) offering improved efficiency and performance in fine-tuning tasks.

Dataset Preparation

- Collected Burmese news across six categories from major news outlets (VOA Burmese, BBC Burmese, RFA).
- Manually annotated by native speakers (April–June 2024).
- Dataset split: 80% training (5.84k sentences) and 20% testing (1.47k sentences).

Table 1: 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%

- Text cleaning and normalization to ensure correct Unicode ordering and reliable segmentation.
- Tokenization:
  - mBERT tokenizer for transformer-based models
  - myWord tool for static embedding models using unigram/bigram dictionaries
- Dataset includes representative examples for all six categories.

Dataset Preparation (Cont'd)

Class	Example Sentences
Sports	အကြို ဗိုလ်လုပွဲ တွင် အနိုင် ရ ရှိ သည့် နှစ် သင်း က ဗိုလ်လုပွဲ ထပ်မံ ယှဉ်ပြိုင် ရ မည် ဖြစ် ပြီး ရှုံးနိမ့် သည့် နှစ် သင်း က တတိယ နေ ရာ လု ပွဲ ဆက်လက် ကစား ရ မည် 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 သူ့ ကို ခြစား မှု ၊ အလွဲသုံးစား မှု နဲ့ အမြတ်ထုတ် မှု စွပ်စွဲ မှု တွေ့ မ ရှိ ဖူး သေး ပါ ။ မစ္စတာ အူမဲရောဗ် ဟာ ဟော့လော့စ် ပါတီ ကနေ ရွေးကောက်ပွဲ ဝင် ရင်း နိုင်ငံ ရေး လောက ထဲ ကို ၂၀၁၉ မှာ ဝင်ရောက် လာ ခဲ့ တယ် ။ အဲဒီ နောက် အစိုးရ အရာရှိ 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	နိုင်ငံတကာ အာကာသ စခန်း ဟာ လာ မယ့် ၂၀၃၀ အထိ အလုပ် လုပ် နေ ဦး မှာ ဖြစ် ပြီး ၂၀၃၁ မှာ ပစိဖိတ် 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 ငါး လ အတွင်း ပြည်ပ သို့ ရေ ထွက် ပစ္စည်း တင် ပို့ မှု မှ ဒေါ်လာ ၂၄၁ သန်း ကျော် ရ ရှိ ပြီး ပင်လယ် ရေ ကြောင်း ကုန်သွယ် မှု မှ ဒေါ်လာ ၁၃၄ သန်း ကျော် တင် ပို့ ထား ကြောင်း စီးပွား ရေး နှင့် ကူးသန်းရောင်းဝယ် ရေး ဝန်ကြီးဌာန မှ သိ 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 လီဆာ အင်စတာဂရမ် တွင် ပို့စ် တစ် ခု တင် လိုက် သည် နှင့် အမေရိကန် ဒေါ်လာ ၅၇၅၀၀၀ ရ နေ ပြီး သား ဟု ဆိုရှယ်မီဒီယာ မားကတ်တင်း ပလက်ဖောင်း က ထုတ်ပြန် သည့် အင်စတာဂရမ် ဖြင့် 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 တွေ့ ကို အမဲ လိုက် သတ်ဖြတ် နေ တာ နဲ့ ပက်သက် ပြီး ဥပဒေ အရ အရေးယူ တာ တွေ့ ၊ မျောက် လေး တွေ့ ကို ထိန်းသိမ်း တာ တွေ့ ကို တော့ မ ကြား မိ သေး ဘူး လို့ ကိုဝင်းပိုင်ဦး 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 2: Example Sentences from Each News Type from News Classification Dataset.

Methodology: Embeddings

- Goal: evaluate different embedding strategies combined with different classification heads while keeping the backbone frozen.
- Neural embeddings map tokens into continuous vectors capturing semantic and syntactic relationships.
- **Static:** TF-IDF, Random and, fastText embeddings.
- **Contextual:** mBERT and Distil-mBERT encoders.
- Static embeddings provide fixed representations; contextual embeddings encode tokens based on sentence context.
- Only the classification head is fine-tuned for all setups.

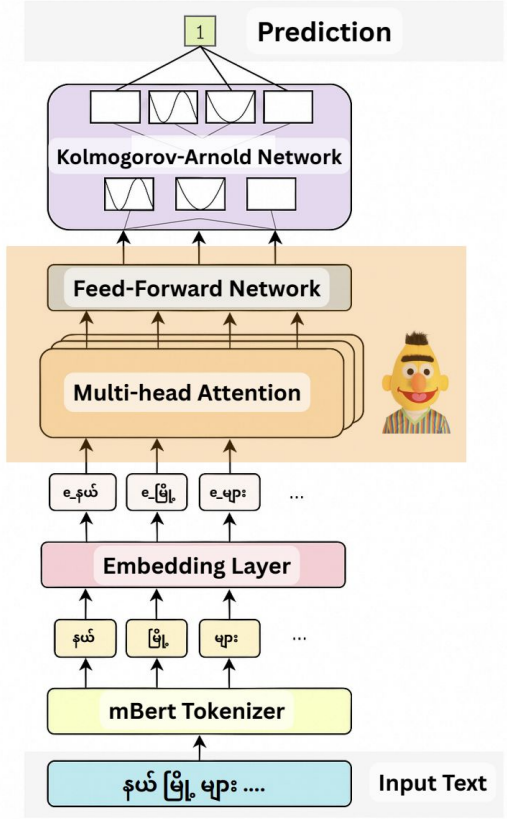


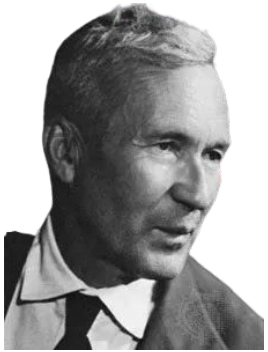
Figure 1: Integration of KAN as a classification head on top of a Transformer encoder for context-aware prediction

Methodology: Kolmogorov-Arnold Representation Theorem

The works of Vladimir Arnold and Andrey Kolmogorov established that if f is a multivariate continuous function, then f **can be expressed** as a finite composition of continuous functions of a single variable and the binary operation of addition.

$$f(\mathbf{x}) = f(x_1, \cdots, x_n) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right),$$

where $\phi_{q,p}: [0,1] \to \mathbb{R}$ and $\Phi_q: \mathbb{R} \to \mathbb{R}$.



Andrey Kolmogorov

Vladimir Arnold

Methodology: Kolmogorov-Arnold Network

KAN: If f is a multivariate continuous function, then f can be **approximated** as a finite composition of continuous functions of a single variable and the binary operation of addition.

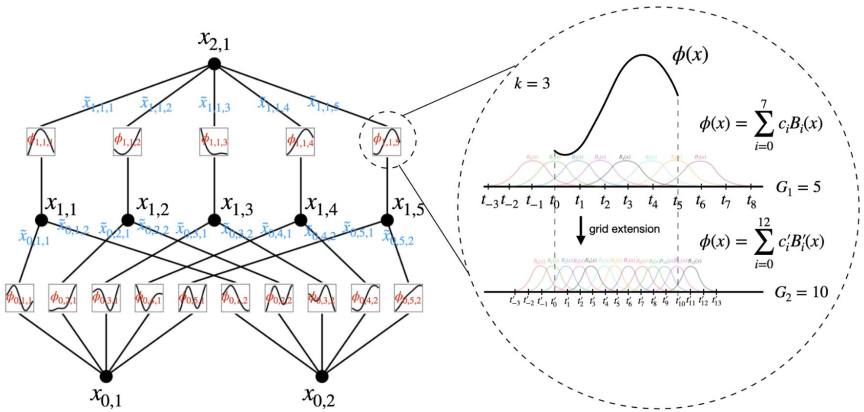


Figure 2.2: Left: Notations of activations that flow through the network. Right: an activation function is parameterized as a B-spline, which allows switching between coarse-grained and fine-grained grids.

Ziming Liu

Max Tegmark

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljacic, M., Hou, T., & Tegmark, M. (2025). KAN: Kolmogorov–Arnold Networks. In Y. Yue, A. Garg, N. Peng, F. Sha, & R. Yu (Eds.), *International Conference on Representation Learning* (pp. 70367–70413).

Methodology: Kolmogorov-Arnold Network (Cont'd)

- **FourierKAN**
  - Replaces spline activations with Fourier series expansions.
  - Approximates functions as sums of sine and cosine terms:

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

  - Efficiently models smooth, periodic, and non-linear patterns.
  - Benefits: smoother gradients, easier initialization, compact parameterization.
- **EfficientKAN**
  - Enforces sparsity via L1 regularization on weights instead of expanded inputs.
  - Optional learnable scaling for activation functions allows trade-off between accuracy and efficiency.
- **FasterKAN**
  - Reduces memory usage by applying activations first, then linear combination.
  - Uses Reflectional Switch Activation Functions (RSWAF) to approximate splines.
  - Benefits: lower memory, faster forward/backward computations, flexible basis functions.

Experimental Setup

- Experiments conducted on Google Colab with NVIDIA Tesla T4 (16 GB VRAM).
- Implemented in PyTorch; evaluation with scikit-learn; pre-trained backbones via Hugging Face Transformers.
- **Embeddings:**
  - Static: TF-IDF, random, fastText
  - Contextual: mBERT, Distil-mBERT
- **Classification heads:**
  - **MLP** – Feed-forward with ReLU
  - **FourierKAN** – Fourier-based, grid size 8
  - **EfficientKAN** – Spline-based, grid size 8, cubic splines
  - **FasterKAN** – Grid-based, learnable grid, inverse denominator formulation

Experimental Setup (Cont'd)

- Optimizer: AdamW with two-tier LR: 2e-4 for head
- Scheduler: Cosine Annealing
- Loss: Cross-entropy
- Training:
  - Static embeddings: 15 epochs, batch size 32
  - Transformer-backbones: 5 epochs, batch size 8
  - Gradient clipping (max-norm = 1.0)
  - Dropout: 0.3; Early stopping with patience = 3
- Evaluation:
  - Weighted F1-score on test set
  - Computational metrics: total/trainable parameters, training time, forward/backward latency

Result and Discussion: Efficiency

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

Table 3: Efficiency comparison of models with different embeddings

Result and Discussion: Efficiency (Cont'd)

- **Training time:**
  - MLP generally fastest across embeddings.
  - **FasterKAN** often trains faster than MLP despite more parameters.
  - **EfficientKAN** requires significantly longer training.
  - Transformer-based models (mBERT, Distil-mBERT) are slowest (hundreds–over 1000 sec).
- **Inference and backward propagation latency:**
  - MLP: lowest latency (forward 0.26–0.46 ms, backward 0.64–0.98 ms).
  - FourierKAN: slightly higher latency but efficient (forward 0.56–1.08 ms, backward 1.23–1.90 ms).
  - FasterKAN: moderate overhead (forward 0.58–0.78 ms, backward 1.0–1.41 ms).
  - EfficientKAN: slowest among non-transformers (forward 1.76–4.49 ms, backward up to 4.18 ms).
  - Transformers: high latency (Distil-mBERT forward 52–100 ms, backward 113–212 ms; mBERT forward 115+ ms, backward 215–418 ms)

Result and Discussion (Cont'd)

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

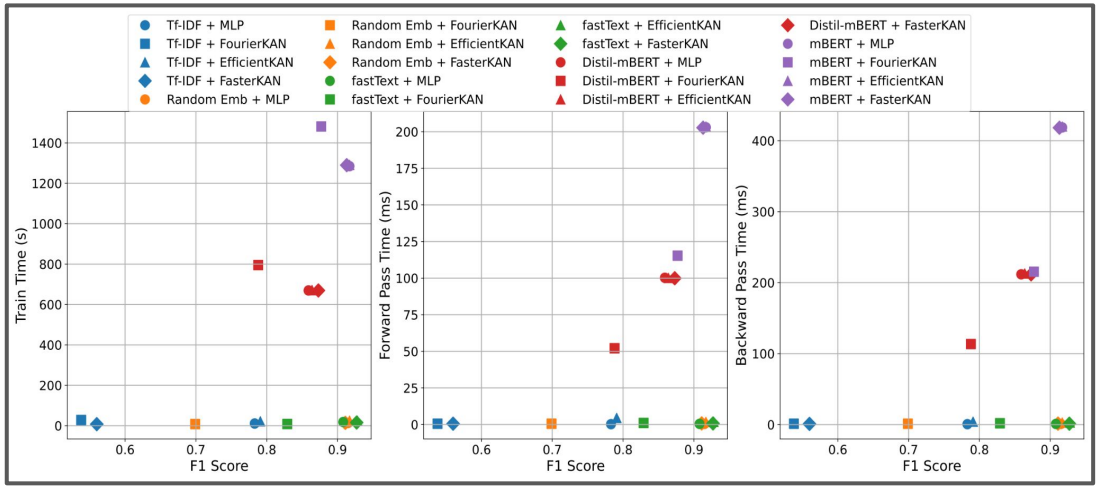


Figure 2: Comparison of model performance versus computational cost

Table 4: F1-score comparison of classification heads across embeddings

Result and Discussion: Performance (Cont'd)

- **Effect of classification head (F1-score):**
  - **EfficientKAN** highest for classical embeddings (e.g., TF-IDF: 0.791).
  - MLP and FasterKAN competitive; FourierKAN generally lower.
  - For transformer embeddings, MLP or FasterKAN perform best; FourierKAN slightly lower but competitive.
- **Effect of embedding type:**
  - Neural Network embeddings (Random, fastText) outperform TF-IDF across all heads.
  - fastText usually yields highest F1, followed by Random and Fourier.
  - Transformer embeddings give modest additional gains over fastText.

Conclusion and Future Work

- Explored KAN variants (EfficientKAN, FasterKAN, FourierKAN) as classification heads for Burmese news sentence classification.
- **Key findings:**
  - Embedding type has the largest impact on performance.
  - Neural Network based word embeddings (Random, fastText) outperform TF-IDF.
  - Transformer embeddings (mBERT, Distil-mBERT) achieve highest accuracy.
  - **EfficientKAN** performs best with static embeddings (fastText: 0.928 F1).
  - **FasterKAN** balances speed and accuracy efficiently.
  - **FourierKAN** underperforms due to less adaptive basis and higher parameter count.
  - For transformer embeddings, head choice matters less; MLP, EfficientKAN, FasterKAN similar.

Conclusion and Future Work (Cont'd)

- Extend KAN-based heads to other low-resource NLP tasks.
- Explore new basis functions for FourierKAN and spline-based KANs.
- Optimize training efficiency for EfficientKAN.
- Release code and dataset for reproducibility and broader adoption.
- Investigate integration with multilingual and cross-lingual embeddings for scalability.

Thank You