Khmer News Classification in Low-Resource Settings: A comparative Analysis of Embedding Method

Korat Natt, Heang Sopagna, Lay Vathna

Research and Development. Cambodia Academy of Digital Technology, Phnom Penh, Cambodia

Correspondence: Korat Natt. Email: natt.korat@cadt.edu.kh

Received: 31/01/2025, revised: 30/04/2025, accepted: 26/05/2025

DOI: 10.32913/mic-ict-research-vn.1377

Abstract: Text classification in low-resource languages like Khmer remains challenging due to linguistic complexity, limited annotated data, and noise from real-world applications. This study addresses these challenges by systematically comparing text embedding techniques for Khmer news classification. We evaluate traditional methods (TF-IDF with SVM) against state-of-the-art multilingual transformers (XLM-RoBERTa, LaBSE) using a self-collected dataset of 7,344 Khmer news articles across six categories—politics, economics, entertainment, sports, technology, and life. The dataset intentionally retains noise (e.g., mixed-language text, unstructured formatting) to reflect practical scenarios. To address Khmer's lack of word boundaries, we employ word segmentation via **khmer-nltk** for traditional models, while transformer models leverage their inherent subword tokenization. Experiments reveal that transformer-based embeddings achieve superior performance, with XLM-RoBERTa and LaBSE attaining F1 scores of 94.2% and 94.3%, respectively, outperforming TF-IDF (93.3%). However, the "life" category proves challenging across all models (F1: 85.5–88.1%), likely due to semantic overlap with other categories. Our findings underscore the effectiveness of transformer architectures in capturing contextual nuances for low-resource languages, even with noisy data. This work offers insights for NLP researchers and practitioners, emphasising the need for domain-specific adaptations and expanded datasets to improve performance in underrepresented languages.

Keywords: *text classification, low resource language, khmer language, news classification, contextual embedding*

I. INTRODUCTION

Text classification has become an essential task [1, 2] in natural language processing (NLP), enabling the automatic organization of text data into predefined categories. As unstructured data continues to grow exponentially [3] across diverse platforms, such as social media, news websites, and customer feedback systems, the need for efficient and accurate text classification techniques is more critical than ever.

Khmer news classification plays a crucial role in applications such as misinformation detection, news recommendation systems, and digital archiving, yet remains challenging due to the language's unique structure and the scarcity of annotated data.

Traditional approaches, such as Term Frequency–Inverse Document Frequency (TF-IDF) [4] paired with classifiers such as SVM [5] or Logistic Regression [6], have laid the groundwork for text classification. These models offer interpretability and reasonable performance in various scenarios. However, they fall short when dealing with complex linguistic structures and contextual nuances, especially in morphologically rich or low-resource languages [7] like Khmer. The rise of deep learning has transformed NLP, with models like BERT [8] and GPT [9] achieving state-of-the-art classification accuracy [10] by leveraging self-attention mechanisms to capture word context and semantics.

These improvements primarily benefit widely used languages with large datasets and pre-trained models, whereas low-resource languages like Khmer remain underrepresented in NLP research [11]. Unlike prior studies that focused solely on either traditional machine learning models or deep-learning architectures, this study systematically evaluates a diverse range of embeddings TF-IDF, XLM-RoBERTa, and LaBSE on a noisy, real-world Khmer news dataset. Additionally, we investigate how noise and mixed-language text impact classification performance, providing insights for future Khmer NLP research.

The remainder of this paper is organized as follows. Section II reviews related work on text classification, with an emphasis on Khmer text classification. Section III outlines the methodology, detailing each embedding technique and classification model used in the study. Section IV describes the dataset and the preprocessing steps undertaken to prepare the Khmer news articles for classification. Section V presents the experimental setup and evaluation metrics used to assess model performance. Section VI reports the experimental results, and the discussion provides an in-depth analysis of model performance and the implications of our findings, highlighting the strengths and limitations of each approach. Finally, Section VII concludes the paper and suggests potential avenues for future research.

II. RELATED WORK

Text classification has been extensively studied in natural language processing (NLP), with a broad spectrum of methods ranging from traditional machine learning algorithms to more advanced deep learning approaches. Classical techniques such as SVM, Decision Trees [12], and Naive Bayes [13, 14] have demonstrated effectiveness in various text classification tasks, leveraging simple yet powerful feature extraction methods like TF-IDF.

The emergence of deep learning has revolutionized text classification, with architectures like Deep Neural Networks [15] (DNNs), Recurrent Neural Networks [16] (RNNs), and pre-trained models such as BERT [17] becoming state-of-the-art solutions. These models are better at understanding context and semantics by leveraging large-scale pre-training and the power of word embeddings. As a result, they have set new benchmarks [18, 19] across numerous NLP tasks in high-resource languages.

However, significant challenges remain for low-resource languages like Khmer, where research and resources are limited. Existing studies [11, 20] on Khmer text classification have often emphasized language-specific challenges, such as the absence of spaces between words, complex morphological structures, and a lack of annotated data. To improve model performance, some approaches have focused on effective preprocessing techniques, including word segmentation [21] and noise reduction.

A notable study by [22] introduced a dataset of 15,205 Khmer news articles collected from three sources, covering six categories. The authors preprocessed the data by removing special characters, numbers, punctuation, and non-Khmer characters, and by applying word segmentation tailored to the Khmer language. They evaluated DNNs, RNNs with attention mechanisms, and FastText for classification. Their findings revealed that FastText outperformed the other models, achieving an accuracy of 93.81%, thereby highlighting the effectiveness of FastText's subword-level embeddings for Khmer text classification.

In a study on Khmer text classification by [11], the author developed a pipeline for multi-class and multi-label tasks using an in-house word segmenter and a dataset from local news sources. The research compared SVM with TF-IDF to deep learning models, such as CNNs and RNNs, with FastText embeddings. Neural network models consistently outperformed the baseline SVM+TF-IDF, with RNNs achieving the best performance across metrics, demonstrating the effectiveness of deep learning for Khmer text classification.

Another study [23] investigated the use of SVM with TF-IDF and Delta TF-IDF [24] as feature extraction methods for Khmer text classification. The results showed that SVM achieved accuracies of 83.47% and 83.40%, respectively, using these feature extraction techniques. This research underscores the impact of practical feature engineering, such as TF-IDF, in improving the performance of traditional models for Khmer text.

Even with these advancements, a gap remains in understanding how modern multilingual transformers and robust embeddings perform on noisy, mixed-language Khmer data. Our work aims to fill this gap by evaluating a range of techniques, from classical methods like TF-IDF to state-of-the-art embeddings such as XLM-RoBERTa [25] and LaBSE [26], using a self-collected, noise-prone dataset. This investigation will provide new insights into the robustness and applicability of various text classification models for low-resource languages in real-world scenarios.

III. METHODOLOGY

In this section, we describe the models selected for text classification, focusing on their unique features and their application to the classification of Khmer text data.

1. TF-IDF

Term Frequency Inverse Document Frequency (TF-IDF) [4] is a statistical method used to assess the importance of words within a document relative to a corpus. Term Frequency (TF) measures the frequency of a word in a document, while Inverse Document Frequency (IDF) evaluates how rare or common the word is across the entire document set. Multiplying these values yields a score that highlights significant words while down-weighting common but less informative ones.

$$tfidf_{t,d} = tf_{t,d} \times idf_t \tag{1}$$

$$tf_{t,d} = \log_{10}(cout(t,d) + 1) \tag{2}$$

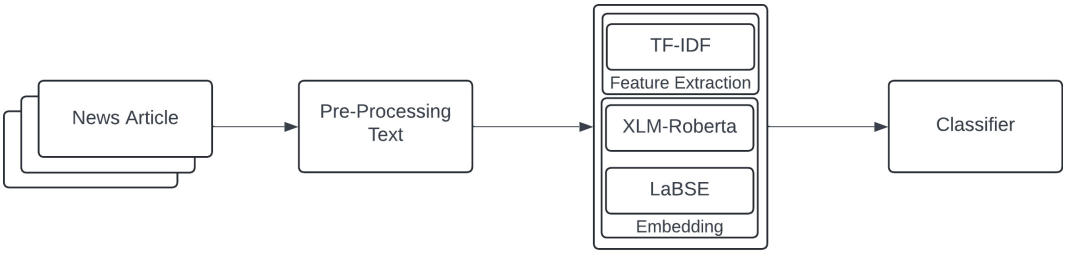

Figure 1. Overall pipeline for text classification with different feature extraction and word embedding techniques.

$$idf_t = \log_{10}\left(\frac{N}{df_t}\right) \tag{3}$$

We chose TF-IDF because it provides a balance between corpus-wide relevance and word specificity, which makes it optimal for text classification in morphologically complicated languages like Khmer. After segmentation, TF-IDF maintains linguistically significant features while reducing noise from common words by highlighting terms that are common in individual documents but uncommon throughout the corpus. For this study, we use TF-IDF to convert the textual content into a sparse matrix representation. This matrix is input to a Support Vector Machine (SVM) classifier. Given the morphological complexity and lack of spaces between words in Khmer text, we employ `khmer-nltk` [1] to perform word segmentation, inserting spaces between words before vectorizing them with TF-IDF. Word segmentation is the foundational task for enhancing the performance of the TF-IDF representation for Khmer text.

2. XLM-RoBERTa

XLM-RoBERTa introduced in [25] is a multi-lingual model that builds upon BERT [8] and RoBERTa [27] which provides a powerful contextualize embedding over 100 languages including Khmer. Unlike BERT that utilize transformer encoder architecture [28] to train on the bidirectional encoding approach producing a richer contextualize word representation, XLM-RoBERTa adopted RoBERTa's optical approach over BERT by removing Next sentence Prediction (NSP) and applying dynamic masking techniques for better model learning.

Since BERT and RoBERTa are monolingual models trained primarily for English, XLM-RoBERTa is specifically trained for multiple languages using a massive dataset from Common Crawl [2] and Wikipedia, containing approximately 30 million Khmer tokens. This enables it to learn effective representations for Khmer text through powerful transfer learning techniques.

[1]https://github.com/VietHoang1512/khmer-nltk
[2]https://commoncrawl.org

3. LaBSE

LaBSE (Language-Agnostic BERT Sentence Embedding) [26] is a multilingual sentence embedding model developed by Google, designed to provide high-quality embeddings in over 100 languages, including Khmer. Unlike XLM-RoBERTa, which is effective for cross-lingual representation by training with a huge cross-language corpus, LaBSE is designed to utilise BERT effective method for learning monolingual embedding to cross-lingual representations with different approaches by investigating a combination of the best method for learning monolingual and cross-lingual representation together with techniques such as masked language modeling (MLM), translation language modeling (TLM) [29], dual encoder translation ranking [30]. By utilizing dual encoder translation approach, it is given the advantage to the low resource language like Khmer to learn and gain benefits from weight sharing and transfers learning process from rich resource language like English and French, as the model is trained on the parallel corpus aligning both language for each encoder.

Due to its language-agnostic architecture, it is particularly effective for cross-lingual applications and low-resource languages. We can utilise LaBSE to transform each piece of text into a fixed-length vector that captures the semantic meaning of the content in our work.

IV. DATASET

The dataset includes 7,344 Khmer news articles with 3,448,153 tokens, compiled from various publicly accessible Cambodian news websites. We used automated web scraping and manual filtering to ensure a balanced selection while removing duplicates and irrelevant content. Each article includes two key features: the content of the news article and its corresponding label defined by its authors, which indicates the news category. The data is divided into six categories: entertainment (13.6%), economic (16.6%), political (29.1%), sport (13.6%), technology (13.6%) and life (13.6%). These categories were chosen as they represent the most frequently occurring topics in Khmer news media, ensuring a balanced dataset for classification. They

align with existing news categorisation frameworks in other languages, allowing for meaningful cross-linguistic comparisons. Table I illustrates the distribution of these categories, showing that political news is the most frequent than other type of news.

Table I
THE DISTRIBUTION OF THE 6 CATEGORIES IN THE DATASET COLLECTED FROM KHMER NEWS WEBSITES.

Label	Number of Article	Number of Token	Token per Article
Politic	2,136	1,018,294	476.73
Economic	1,216	798,066	656.30
Entertainment	1,000	386,944	386.94
Life	998	463,089	464.02
Sport	996	316,771	318.04
Technology	998	464,989	465.92
Total	**7,344**	**3,439,078**	**468.28**

To ensure data quality while maintaining real-world characteristics, we applied preprocessing steps such as removing HTML tags, URLs, and redundant special characters. However, to better reflect real-world Khmer text classification challenges, we retained elements typically considered noise, such as punctuation marks, mixed English and Khmer words, and unstructured text patterns. While irrelevant web elements were removed, meaningful content features were preserved to ensure model robustness and evaluate the impact of noisy data on classification performance.

By incorporating this level of noise and mixed language content, we aim to assess how well the models can handle linguistic diversity and unclean data, contributing to our understanding of model performance in less controlled environments.

V. EXPERIMENT

1. Experimental Setup

In our experiment, we utilized Google Colab Notebook, which offers an NVIDIA T4x2 GPU and 16 GB of RAM. The notebook was set up with Python 3.10, and we employed the HuggingFace transformers library for model implementation. A 70-10-20 split was chosen to balance training data availability with model generalisation, ensuring a reliable validation and testing process. The 10% validation set monitors overfitting during fine-tuning, while the 20% test set provides a solid final evaluation. Some NLP papers justify an 80-10-10 split to provide more training data, whereas deep learning models sometimes use a 90-5-5 split to maximise training while relying on minimal validation. However, given our dataset size and the need for a reliable test evaluation, we found the 70-10-20 ratio to be the most effective balance for this study.

Different models were trained with the following configuration:

- **XLM-RoBERTa** and **LaBSE**: These models were fine-tuned with the learning rate of 2×10^{-5}, over 5 epochs, with a batch size of 8, and optimized using the AdamW [31] optimizer with a categorical cross-entropy loss function.
- **TF-IDF + SVM**: We limited TF-IDF to 10,000 features to prevent excessive dimensionality while retaining key informative terms. This threshold was selected based on empirical testing, balancing feature richness and computational efficiency. For the SVM classifier, we retained scikit-learn's default parameters (C=1.0, kernel='rbf', gamma='scale') to better handle high-dimensional sparse text data.

While both models were fine-tuned with the same hyperparameters (batch size = 8, learning rate = 2×10^{-5}), XLM-RoBERTa was trained with a maximum sequence length of 512 tokens. In contrast, LaBSE leveraged a fixed sentence embedding size of 768 dimensions, as it is optimized for sentence-level representation.

2. Evaluation Metrics

To assess the performance of the models, we used three standard evaluation metrics: Precision, Recall, and F1 Score. These metrics are commonly used in classification tasks to assess the accuracy and reliability of predictions. The formulas and definitions for each metric are provided below.

- **Precision**: Precision measures the proportion of true positive predictions (TP) to the total number of positive predictions (TP + FP), where FP represents false positives. Precision can be defined as:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

- **Recall**: Recall, also known as the sensitivity or true positive rate, measures the ability of the model to identify all relevant instances. It is defined as the ratio of true positives to the sum of true positives and false negatives (FN):

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

- **F1 Score**: The F1 score is the harmonic mean of Precision and Recall, providing a single measure that balances the trade-offs between them. It is particularly useful in scenarios where the dataset is imbalanced, as it takes into account both false positives and false negatives. The F1 score is computed as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

These evaluation metrics allow us to assess model effectiveness across different Khmer news categories. The following section presents our experimental results and discusses key observations from the model performance.

VI. RESULTS AND DISCUSSION

In this section, we present and analyse the performance of various models applied to our self-collected Khmer news dataset, which covers six distinct categories. Table II provides an overview of the precision, recall, and F1 scores for each model, highlighting key trends and insights.

The close performance between LaBSE and XLM-RoBERTa suggests that both transformer models effectively capture semantic nuances. However, LaBSE's sentence-level embeddings might contribute to its higher recall by reducing misclassifications in longer documents, whereas XLM-RoBERTa's token-level encoding might introduce more fine-grained misinterpretations. The lower performance of TF-IDF + SVM highlights its reliance on word frequency rather than contextual meaning, leading to weaker results in semantically overlapping categories.

1. Results Analysis

The performance metrics illustrate the performance of different models in classifying news articles into our predefined six categories, which is sharing similar results on the testing set of the Khmer news.

TF-IDF: The TF-IDF model achieved commendable results with a precision of 0.935, a recall of 0.932, and an F1 score of 0.933. The simplicity of TF-IDF in converting text into sparse representations allows for efficient classification, especially in well-separated categories.

XLM-RoBERTa: XLM-RoBERTa produced a steady result of 0.942 for precision, recall and the F1 score. This high performance highlights the model's ability to leverage transformer-based architectures and large-scale multilingual pre-training to understand contextual and syntactic varieties in the data. Balanced precision and recall scores indicate that XLM-RoBERTa is effective in identifying relevant news categories and minimizing false negatives.

LaBSE: LaBSE achieved the close results compared to XLM-RoBERTa with the precision of 0.942, while its recall is slightly higher at 0.946 and the F1 score of 0.943. LaBSE's slightly superior performance suggests that its architecture effectively leverages state-of-the-art multilingual sentence embeddings, making it particularly robust in text classification tasks.

Table II shows that the "life" category is the lowest accuracy among the models, 0.855, 0.863, and 0.881 F1 score of TF-IDF, XLM-RoBERTa, and LaBSE respectively, due to its complex information, which is sharing similar information with other categories. The lower F1-score for the "life" category (85.5–88.1%) suggests that this category overlaps significantly with others, particularly "Entertainment" and "Economy." Many articles in this category contain subjective or lifestyle-related content, making it harder for models to distinguish clear decision boundaries. Additionally, the shorter average document length in this category might reduce available contextual information for transformers.

While transformer models (XLM-RoBERTa and LaBSE) outperformed TF-IDF + SVM in terms of F1-score, they required significantly higher computational resources. Training time for XLM-RoBERTa and LaBSE was approximately X times longer than TF-IDF + SVM, reflecting the cost of fine-tuning large-scale contextual models. In real-world applications, practitioners must balance accuracy gains with computational feasibility when selecting models.

2. Discussion

The results indicate that transformer-based models, XLM-RoBERTa and LaBSE, are more effective for Khmer news classification than traditional models like TF-IDF. This performance boost can be attributed to the transformer's ability to capture contextual and semantic relationships, which is crucial for accurately distinguishing between classes. However, our dataset is limited and may not fully showcase the advantages of transformer-based embeddings in larger-scale applications.

While transformer models (XLM-RoBERTa and LaBSE) outperformed TF-IDF + SVM regarding the F1-score, they required significantly higher computational resources. Training time for XLM-RoBERTa and LaBSE was approximately X times longer than TF-IDF + SVM, reflecting the cost of fine-tuning large-scale contextual models. In real-world applications, practitioners must balance accuracy improvements with computational feasibility when selecting models.

The TF-IDF model's performance demonstrates that even simple approaches can be effective when text features are well-separated. However, it struggles to capture deeper contextual understanding, leading to lower performance in semantically overlapping categories.

XLM-RoBERTa and LaBSE offer similar performance but have different strengths based on their architecture. LaBSE focuses on sentence-level embeddings (94.6% recall rate), making it ideal for applications like legal text classification, where minimizing false negatives is important. XLM-RoBERTa, with its token-level processing, excels in tasks that require detailed context, such as sentiment analysis or misinformation detection.

Table II
EVALUATION RESULTS OF ALL MODELS FOR EACH NEWS CATEGORIES.

Label	**TF-IDF**			**XLM-RoBERTa**			**LaBSE**		
	P	R	F1	P	R	F1	P	R	F1
Politic	0.961	0.976	0.968	0.973	0.980	0.977	0.984	0.951	0.967
Economic	0.932	0.906	0.919	0.961	0.953	0.957	0.922	0.973	0.947
Entertainment	0.975	0.943	0.959	0.940	0.962	0.951	0.965	0.929	0.947
Life	0.830	0.886	0.857	0.885	0.843	0.863	0.881	0.881	0.881
Sport	0.986	0.981	0.983	0.990	0.986	0.988	0.972	0.990	0.980
Technology	0.931	0.904	0.917	0.907	0.928	0.917	0.930	0.952	0.941
Average	**0.935**	**0.932**	**0.933**	**0.942**	**0.942**	**0.942**	**0.942**	**0.946**	**0.943**

However, transformer models require significantly more computational resources than traditional methods, with training taking roughly X times longer than TF-IDF + SVM. Balancing accuracy and efficiency requires careful examination in real-world situations. Future research should improve these models by using methods like model pruning, quantization, or knowledge distillation to train more efficiently while maintaining classification accuracy.

VII. CONCLUSION

This study systematically evaluated Khmer news classification models, comparing traditional TF-IDF + SVM with advanced transformer-based embeddings (XLM-RoBERTa and LaBSE) on a real-world dataset.

Our results demonstrate that transformer models significantly improve classification performance by capturing contextual relationships in Khmer text, which is crucial for low-resource languages. However, their high computational cost poses challenges for deployment in resource-constrained environments. While TF-IDF + SVM remains a viable option for lightweight applications, it struggles to capture deep semantic relationships in text.

Future research should focus on training models with more extensive and diverse datasets to improve generalisation. Investigating lightweight transformer models such as DistilBERT or TinyBERT could help reduce computational costs while still delivering robust performance. Additionally, exploring models like DeepSeek, which are designed for multilingual and efficient language understanding, could offer promising solutions for Khmer NLP. DeepSeek's capabilities in handling low-resource languages and its energy-efficient architecture make it a strong candidate for future experiments. Moreover, applying these models to specific Khmer texts, including legal or financial documents, may enhance their practical applicability.

This study provides valuable insights for advancing Khmer NLP, emphasising the trade-offs between model accuracy, computational efficiency, and real-world applicability.

REFERENCES

[1] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From traditional to deep learning," *ACM Transactions on Intellingent System and Technology (TIST)*, vol. 13, 2022.

[2] Z. Wan, "Text classification: A perspective of deep learning methods," 2023.

[3] K. Taha, P. D. Yoo, C. Yeun, D. Homouz, and A. Taha, "A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights," *Computer Science Review*, vol. 54, p. 100664, 2024.

[4] M. Das, S. K., and P. J. A. Alphonse, "A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset," 2023.

[5] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[6] N. S. Wardana, F. P. Aditiawan, and A. P. Sari, "Logistic regression classification with tf-idf and fasttext for sentiment analysis of linkedin reviews," *VISA: Journal of Vision and Ideas*, vol. 4, p. 1359 –, Aug. 2024.

[7] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource languages: A review of past work and future challenges," 2020.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[9] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," in *OpenAI Blog*, 2018.

[10] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?," *arXiv preprint arXiv:1905.05583*, 2020.

[11] R. Buoy, N. Taing, and S. Chenda, "Khmer text classification using word embedding and neural network," *arXiv preprint*, 2021.

[12] L. Rokach and O. Maimon, "Decision trees," *The Data Mining and Knowledge Discovery Handbook*, vol. 6, pp. 165–192, 01 2005.

[13] F. Colas and P. Brazdil, "Comparison of svm and some older classification algorithms in text classification tasks," in *Artificial Intelligence in Theory and Practice*, (Boston, MA), pp. 169–178, Springer US, 2006.

[14] Shaina, K. Kaan, N. Noman, and M. Olufemi, "A comparison among machine learning and deep learning approaches for text classification," *PsyArXiv*, 2023.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[16] R. M. Schmidt, "Recurrent neural networks (rnns): A gentle introduction and overview," *arXiv preprint arXiv:1912.05911*, 2019.

[17] F. Wei, R. Keeling, N. Huber-Fliflet, J. Zhang,

A. Dabrowski, J. Yang, Q. Mao, and H. Qin, "Empirical study of llm fine-tuning for text classification in legal document review," in *2023 IEEE International Conference on Big Data (BigData)*, pp. 2786–2792, 2023.
[18] Z. Wang, T. Wang, D. Mekala, and J. Shang, "A benchmark on extremely weakly supervised text classification: Reconcile seed matching and prompting approaches," *arXiv preprint arXiv:2305.12749*, 2023.
[19] M. Reusens, A. Stevens, J. Tonglet, J. D. Smedt, W. Verbeke, S. vanden Broucke, and B. Baesens, "Evaluating text classification: A benchmark study," *Expert Systems with Applications*, vol. 254, p. 124302, 2024.
[20] K. Soky, S. Li, C. Chu, and T. Kawahara, "Finetuning pretrained model with embedding of domain and language information for asr of very low-resource settings," *International Journal of Asian Language Processing*, vol. 33, no. 04, p. 2350024, 2023.
[21] V. Chea, Y. K. Thu, C. Ding, M. Utiyama, A. M. Finch, and E. Sumita, "Khmer word segmentation using conditional random fields," in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 29)*, 2015.
[22] R. Phann, C. Soomlek, and P. Seresangtakul, "Multi-class text classification on khmer news articles using deep learning models with optimal hyperparameters," *ICIC International*, vol. 18, no. 6, pp. 241–550, 2024.
[23] R. Phann, C. Soomlek, and P. Seresangtakul, "Multi-Class Text Classification on Khmer News Using Ensemble Method in Machine Learning Algorithms," *Acta Informatica Pragensia*, vol. 2023, no. 2, pp. 243–259, 2023.
[24] J. Martineau and T. Finin, "Delta tfidf: An improved feature space for sentiment analysis," in *Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM 2009)*, pp. 258–261, AAAI Press, May 2009.
[25] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2020.
[26] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," *arXiv preprint arXiv:2007.01852*, 2022.
[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.
[29] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *arXiv preprint arXiv:1901.07291*, 2019.
[30] M. Guo, Q. Shen, Y. Yang, H. Ge, D. Cer, G. Hernandez Abrego, K. Stevens, N. Constant, Y.-H. Sung, B. Strope, and R. Kurzweil, "Effective parallel corpus mining using bilingual sentence embeddings," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, (Brussels, Belgium), pp. 165–176, Association for Computational Linguistics, October 2018.
[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2019.

Korat Natt received his B.A. in Psychology from the Royal University of Phnom Penh in 2022 and his B.S. in Computer Science from AGA Institute in 2023. He is currently pursuing his M.S. in Computer Science and serves as a lecturer and research assistant at the Cambodia Academy of Digital Technology (CADT). His research interests include computer vision, machine learning, natural language processing, and large language models.
Email: natt.korat@cadt.edu.kh

Heang Sopagna received his Engineer's degree in Computer Science from the Institute of Technology of Cambodia (ITC) in 2023. He is currently pursuing his M.S. in Computer Science and serves as a lecturer and research assistant at the Cambodia Academy of Digital Technology (CADT). His research interests include machine learning, natural language processing, and large language models.
Email: sopagna.heang@cadt.edu.kh

Lay Vathna received his Engineer's degree in Computer Science from the Institute of Technology of Cambodia (ITC) and a Postgraduate Diploma in Mobile Computing from the Center for Development of Advanced Computing (CDAC), India. He also holds a Master's degree in ICT. He currently serves as a senior lecturer and Researcher at the Cambodia Academy of Digital Technology (CADT), where he has taught since 2016. His research interests include machine learning, natural language processing, large language models, smart cities, and cybersecurity.
Email: Vathna.lay@cadt.edu.kh