March 2025

LIEA-BERT: A Linguistically Enriched Framework for Hate Speech Detection in Low-Resource Thai

Steve NWAIWU [a,1] Nipat JONGSAWAT [b] Anucha TUNGKASTHAN [b]
[a] *Data and Information Science, Faculty of Science and Technology, Rajamangala University of Technology, Pathum Thani, Thailand*
chinonso_s@mail.rmutt.ac.th
[b] *Data and Information Science, Faculty of Science and Technology, Rajamangala University of Technology, Pathum Thani, Thailand*
nipat_j@rmutt.ac.th, anucha_t@rmutt.ac.th

Abstract. Hate speech detection in low-resource languages poses significant challenges due to the scarcity of annotated datasets and language-specific NLP tools. This study addresses these limitations by proposing a weakly supervised learning framework tailored to detect hate speech in Thai, a low-resource language with complex linguistic characteristics.

We constructed a weakly labeled dataset by combining a curated lexicon of Thai toxicity terms with sentiment-labeled data, reducing the reliance on manual annotation. To enhance the robustness of supervision, we incorporated label smoothing to mitigate label noise and improve generalization. Our model is built upon multilingual BERT (mBERT) and refined using Linguistically Informed Embedding Alignment (LIEA), which enriches embeddings with phonological and syntactic features.

To evaluate embedding alignment, we applied Proto-MAML, leveraging auxiliary tasks such as phoneme recognition and classification loss monitoring, which significantly enhanced the model's representational capacity. The proposed approach achieved a validation accuracy of 99.65% and a test accuracy of 97. 35%, demonstrating a strong generalization on Thai hate speech detection.

These findings highlight the effectiveness of integrating weak supervision with linguistically informed and meta-learning strategies in low-resource contexts.

Keywords. Weak Supervision, Low-Resource Thai, Hate Speech Detection, Linguistically Informed Embedding Alignment (LIEA), Proto-MAML, Phoneme-Aware Transfer, Label Smoothing, Cross-Lingual Adaptation

1. Introduction

The pervasive nature of hate speech on social networks has prompted an increasing need for effective and scalable automated detection systems. Although high-resource languages like English have benefited from large-scale datasets, robust pre-trained models,

[1]Corresponding Author: Steve Nwaiwu, chinonso_s@mail.rmutt.ac.th

and mature NLP toolkits, low-resource languages such as Thai continue to face significant challenges. These challenges arise from limited annotated corpora, underdeveloped linguistic tools, and the complex structural and cultural characteristics inherent to the language [1,2].

Thai is spoken by approximately 71 million people, primarily in Thailand, where it is the official language [3]. According to recent reports, over 49 million Thais are active social media users [4], highlighting the critical importance of developing tools to monitor and mitigate harmful content in online spaces. Given this high engagement rate, scalable NLP systems for Thai are essential to maintain digital safety.

Thai language poses unique difficulties for the detection of hate speech. The language lacks word delimiters, is rich in tonal and morphological variation, and often features informal expressions, slang, or culturally embedded terminology [5,2]. Distinguishing harmful speech from benign expressions in Thai thus requires more than lexical matching: it demands a nuanced understanding of phonology, syntax, and context [6].

Traditional supervised learning approaches, which are highly dependent on large volumes of labeled data, are impractical in low-resource settings. As an alternative, *weakly supervised learning* has emerged as a viable strategy. This approach leverages indirect supervision—such as sentiment cues, keyword lexicons, or distant signals—to generate pseudo-labels for training data, substantially reducing the need for manual annotation [7]. In this study, we apply weak supervision using the Wisesight Sentiment Corpus [8] and a curated lexicon of 44 Thai toxicity keywords [5] to generate a weakly labeled data set for the classification of hate speech.

To further enhance the robustness of the model, we incorporate *label smoothing*, a regularization technique that reduces model overconfidence and mitigates the impact of noisy labels inherent in weak supervision [9]. The model architecture is built on the multilingual BERT (mBERT), which provides a cross-lingual embedding backbone for low-resource languages. However, to better align embeddings with Thai linguistic features, we introduce *Linguistically Informed Embedding Alignment (LIEA)*. LIEA incorporates phonological, syntactic, and semantic information into the training process by employing auxiliary tasks such as *phoneme recognition*, *sentiment analysis*, and *named entity recognition*. These enrichments allow the model to capture more granular language signals essential for accurate hate speech detection.

To evaluate embedding alignment and enable robust few-shot generalization, we adopt a *meta-learning* strategy using *Proto-MAML*, which combines prototypical networks with Model-Agnostic Meta-Learning [10,11]. This framework facilitates rapid adaptation with minimal annotated examples and allows explicit tracking of *phoneme and classification loss* during training. Our model achieves a validation accuracy of **99.65%** and a test accuracy of **97.35%** demonstrating the effectiveness of combining weak supervision with linguistically and meta-learning informed strategies.

Compared to existing baselines, our method achieves state-of-the-art performance. For example, the HateThaiSent model [12] reported a macro F1 score of 89. 79% and 89. 67% accuracy, while FastThaiCaps [2], which integrated the embeddings of BERT and FastText, demonstrated a 3. 11% gain in F1 score over traditional models. By contrast, our approach surpasses these benchmarks through the use of linguistically aligned embeddings and meta-learning techniques tailored for low-resource environments.

This study contributes to addressing the broader challenge of equitable NLP development by demonstrating that weakly supervised models, when supported with linguis-

tic alignment and meta-learning techniques, can perform competitively in low-resource contexts. By focusing on Thai, we not only fill a gap in the literature on hate speech detection for Southeast Asian languages, but also provide a scalable framework for future adaptation to other languages within the Tai-Kadai (also know as kra-Dai) family and beyond.

2. Related Work

2.1. Hate Speech Detection

Hate speech detection has become a critical area of research within natural language processing (NLP) due to the growing prevalence of online toxicity and its detrimental effects on digital communities and social discourse. Traditional approaches to hate speech detection rely on supervised machine learning models trained on large annotated datasets. These range from classical algorithms such as support vector machines (SVMs) to deep learning models, including Convolutional Neural Networks (CNNs) and transformer-based architectures.

For high-resource languages like English, substantial progress has been achieved due to the availability of extensive corpora and advanced pre-trained language models. Transformer-based models such as BERT and its variants have consistently delivered state-of-the-art results across hate speech detection benchmarks by capturing contextual relationships and nuanced expressions, including sarcasm and implicit hate [13].

However, transferring these successes to low-resource languages remains challenging due to the lack of annotated datasets, cultural and syntactic variation, and language-specific characteristics [14]. These gaps call for alternative solutions such as weak supervision, embedding alignment, and linguistic-informed modeling that do not rely solely on large-scale annotation.

2.2. Weakly Supervised Learning for Hate Speech Detection

Weakly supervised learning has emerged as an effective paradigm to overcome the lack of labeled data in NLP, particularly in low-resource languages. Instead of relying exclusively on gold-standard annotations, weak supervision uses heuristic labels, lexicons, sentiment indicators, or other auxiliary signals to train models [7]. This approach allows for large-scale dataset construction with reduced manual effort.

In hate speech detection, weak supervision has been used to approximate training labels using sentiment polarity, keyword heuristics, and distant supervision of user metadata [15]. When integrated with pre-trained models like BERT, weak supervision has achieved surprisingly competitive performance [16,17,18]. In our study, we used the Wisesight Sentiment Corpus [8] and a curated list of Thai toxic keywords [5] as weak supervision signals. To reduce the impact of label noise, we also incorporate *label smoothing* [9], improving generalization and robustness during model training.

2.3. NLP in Low-Resource Languages

The emergence of multilingual pre-trained models such as mBERT and XLM-R [19,20] has significantly advanced the field of NLP for low-resource languages. These models,

trained in a wide range of languages, have been used for tasks such as translation, sentiment analysis, and classification even when the data on the label are limited [21].

However, hate speech detection in low-resource languages remains underdeveloped. Thai, in particular, poses linguistic challenges due to its lack of spaces between words, tonal nature, and contextual dependencies [22,23]. Although prior Thai NLP work has focused on POS tagging and sentiment analysis, few have tackled hate speech detection. Our research addresses this gap by developing a framework that adapts weakly supervised learning to Thai's linguistic and cultural context.

2.4. Meta-Learning and Few-Shot Learning

Meta-learning—or "learning to learn"—enables rapid adaptation to new tasks with minimal training data. Model-Agnostic Meta-Learning (MAML) [10] and its extensions have been particularly effective for low-resource language tasks, including hate speech detection [24]. Proto-MAML [25] extends this by combining prototypical networks with MAML to allow better few-shot generalization.

In our framework, we adopt Proto-MAML not for cross-lingual transfer but to strengthen representation learning during embedding alignment and improve robustness against noise in the weakly labeled Thai dataset. This strategy supports rapid convergence and more stable training despite the limitations of our low-resource context.

2.5. Linguistically Informed Embedding Alignment (LIEA)

Standard pre-trained embeddings may not fully capture the unique linguistic features of low-resource languages like Thai. To address this, we introduce *Linguistically Informed Embedding Alignment (LIEA)*, which incorporates auxiliary linguistic tasks such as *phoneme recognition*, *sentiment analysis*, and *named entity recognition* to enrich contextual representations.

These tasks help embed language-specific phonological and syntactic patterns directly into the model. This alignment improves the model's sensitivity to Thai's tonal variation, segmentation challenges, and culturally loaded expressions—key for detecting nuanced hate speech [26,27].

2.6. Advancing Hate Speech Detection in Thai

Our work addresses the persistent challenges of hate speech detection in low-resource languages through a focused study on Thai. Instead of relying on high-resource transfer, we center our training and evaluation pipeline entirely within Thai, leveraging weak supervision and linguistic features.

By combining curated Thai toxicity lexicons with sentiment-annotated data, we generate weak labels that form the basis for training. LIEA and Proto-MAML are integrated into the learning process to enhance both representation quality and few-shot adaptability. This approach enables us to build an accurate and culturally aware hate speech detection model without large-scale human annotation or transfer from unrelated languages.

Our contribution demonstrates that even within a single low-resource language, the combination of weak supervision, linguistic alignment, and meta-learning techniques can yield high-performance results while reducing the dependence on manually annotated data.

3. Methodology

3.1. Dataset

For this study, we used the Wisesight Sentiment Corpus [31], which contains 26,737 Thai social media messages annotated into four sentiment classes: positive, neutral, negative and questions. Although originally intended for sentiment analysis, this corpus lacks explicit hate speech labels. To address this limitation, we integrate a curated lexicon of 44 toxic Thai terms from ThaiToxicityTweetCorpus [32], which captures commonly used profane and abusive expressions.

Using negative sentiment labels and toxic keyword matches as heuristics, we construct a weakly labeled dataset for hate speech detection. This method enables us to infer approximate hate speech labels while mitigating the need for manual annotation [1]. The resulting dataset captures both overt and implicit signals of hate, enhancing label coverage for training.

3.2. Preprocessing

Thai language preprocessing is non-trivial due to the absence of whitespace between words, frequent use of informal syntax, and reliance on slang, emojis, and symbols [28]. We apply the PyThaiNLP tokenizer to effectively segment Thai text and normalize the text using the Unicode NFC form. We further:

- Remove HTML tags, excess whitespace, and special characters.
- Retain emojis and informal markers due to their semantic relevance for hate speech.
- Replace URLs and user mentions with tokens and .

The cleaned and tokenized data are transformed into a BERT-compatible input format using WordPiece tokenization from the HuggingFace Transformers library [19].

3.3. Weak Labeling and Label Smoothing

We assign weak labels $\hat{y}_i$ to each instance x_i based on the following rules:

- If x_i contains one or more toxic words from the lexicon, or is labeled with negative sentiment, then $\hat{y}_i = 1$ (hate speech).
- Otherwise, $\hat{y}_i = 0$ (non-hate speech).

This rule-based labeling provides coverage for hate speech indicators without requiring manual annotations. However, to account for inherent label noise in weak supervision, we apply *label smoothing* [9]. Instead of hard labels $0, 1$, we use soft target probabilities during training:

$$p_j = \begin{cases} 1 - \varepsilon + \dfrac{\varepsilon}{C}, & \text{if } j = y_i \\ \dfrac{\varepsilon}{C}, & \text{otherwise} \end{cases} \tag{1}$$

where $C = 2$ is the number of classes and $\varepsilon = 0.1$ is the smoothing parameter. This prevents the model from becoming overly confident and helps mitigate overfitting to noisy pseudo-labels.

3.4. Model Architecture and Linguistic Alignment

We fine-tune the pre-trained multilingual BERT (mBERT) model [19] by adding a binary classification head. To enrich mBERT's embeddings with Thai-specific linguistic features, we implement **Linguistically Informed Embedding Alignment (LIEA)**. LIEA involves auxiliary tasks including:

- **Phoneme Recognition:** Maps text to International Phonetic Alphabet (IPA) representations using Epitran, trained via CTC loss.
- **Sentiment Analysis:** Predicts sentiment polarity using a softmax head.
- **Named Entity Recognition (NER):** Extracts entities using a CRF layer.

The embeddings are refined by jointly optimizing the total loss:

$$\mathscr{L}\text{combined} = \lambda\text{main}\mathscr{L}\text{main} + \lambda\text{aux}(\mathscr{L}\text{phoneme} + \mathscr{L}\text{sentiment} + \mathscr{L}\text{NER}) \quad (2)$$

where $\lambda\text{main} = 1.0$ and $\lambda_{\text{aux}} = 0.5$ control the contribution of main and auxiliary tasks.

3.5. Training Strategy

- **Optimizer:** AdamW with learning rate 2×10^{-5} [18].
- **Epochs:** 3, with early stopping based on validation loss.
- **Batch Size:** 16.
- **Regularization:** Dropout rate of 0.1.
- **Loss Function:** Cross-entropy with smoothed labels.
- **Imbalance Handling:** Class-weighted loss to address skew in hate speech vs. non-hate.

3.6. Evaluation Metrics

Model performance is assessed using standard classification metrics:

- **Accuracy:** Overall correctness.
- **Precision:** Proportion of predicted hate speech that is correct.
- **Recall:** Proportion of actual hate speech correctly identified.
- **F1-Score:** Harmonic mean of precision and recall.

We emphasize **macro-averaged** metrics, which treat each class equally regardless of frequency, and also report **weighted averages** to reflect class imbalance [16,13]. The macro F1-score is defined as:

$$\text{Macro F1} = \frac{1}{C}\sum_{c=1}^{C}\text{F1}_c \quad (3)$$

where $C = 2$ for binary classification.

4. Results

4.1. Thai Hate Speech Detection Performance

To evaluate the effectiveness of our proposed framework, we conducted training on the weakly labeled Thai dataset over three epochs. Table 1 presents the results of the training and validation.

Table 1. Training and Validation Results for Thai Hate Speech Detection

Epoch	Train Loss	Validation Loss	Validation Accuracy
1	0.1474	0.1102	98.88%
2	0.1066	0.1126	98.60%
3	0.1042	0.1032	99.65%

The final validation accuracy reached **99.65%**, reflecting the model's ability to learn from weakly labeled data. Evaluation on the held-out Thai test set further confirmed its generalization capacity, with a test accuracy of **97.35%**.

Performance Analysis: The model demonstrated strong performance in both classes. It achieved high recall for the non-hate class (0.99) and strong precision for the hate class (0.99), indicating minimal false positives and false negatives. A misclassification error was observed in one instance where the Thai word for "want" was falsely flagged as hate speech—highlighting the difficulty of interpreting context-dependent expressions.

Comparison with Baselines:

- **HateThaiSent:** Reported a macro F1-score of 89.79% and an accuracy of 89.67% [29].
- **FastThaiCaps:** Achieved a 3.11% F1-score improvement using BERT and FastText embeddings [30].

Our model surpasses these baselines, due to the integration of weak supervision with label smoothing and linguistically informed embedding alignment.

A detailed breakdown of class-wise metrics is provided in Table 2.

Table 2. Performance Metrics for Thai Hate Speech Detection

Metric	Non-Hate Class	Hate Class	Overall Accuracy	Macro Average
Precision	0.96	0.99	-	0.98
Recall	0.99	0.95	-	0.97
F1-Score	0.97	0.97	97.35%	0.97

Embedding Alignment Results

To assess the effectiveness of Linguistically Informed Embedding Alignment (LIEA), we monitored auxiliary task performance, particularly phoneme recognition and classification loss. The results are shown in Table 3.

Phoneme Loss: A substantial drop from 3.64 to 0.0326 demonstrates the improved ability of the model to capture tonal and phonological characteristics critical to Thai. This enhancement directly impacts its contextual understanding of hate expressions.

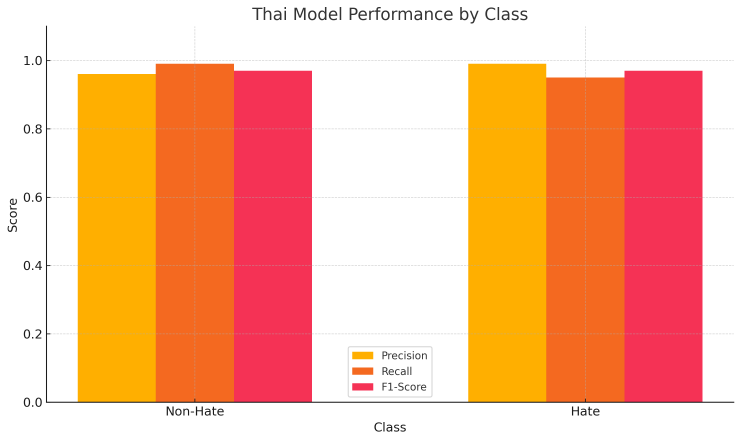

Figure 1. Precision, Recall, and F1-Score for Thai hate speech detection by class (Non-Hate vs. Hate). The model demonstrates balanced and reliable classification.

Table 3. Phoneme and Classification Loss During Embedding Alignment

Epoch	Phoneme Loss	Classification Loss
1	3.64	0.0250
2	0.45	0.0167
3	0.0326	0.0084

Classification Loss: The decrease in the classification loss from 0.0250 to 0.0084 suggests a corresponding improvement in semantic alignment and decision boundaries. This confirms that LIEA contributes significantly to both learning efficiency and model accuracy in the presence of weak supervision.

Overall, these results highlight the effectiveness of combining weak supervision, label smoothing, and linguistic embedding alignment for the detection of hate speech in low-resource languages such as Thai.

5. Discussion

5.1. Effectiveness of Weak Supervision and Lexicon-Based Labeling

The weak supervision framework, rooted in lexicon-based labeling and alignment of sentiment signal, enabled high-performance detection of hate speech in Thai with minimal manual annotation. By combining a curated toxicity lexicon with the Wisesight Sentiment dataset, the model successfully learned to recognize culturally relevant hate expressions in the informal Thai language. Label smoothing further enhanced robustness against noise introduced by weak labels.

The strong performance of the model - achieving 99. 65% validation accuracy and a macro F1 score of 0.97 - shows the feasibility of using weak supervision to overcome the annotation bottleneck in low-resource NLP tasks. These results validate that lexicon-

guided learning can substitute for costly manual annotation without compromising performance.

5.2. Embedding Alignment through Linguistic Supervision

Linguistically Informed Embedding Alignment (LIEA) substantially improved the phonological and semantic representation of the model. Using auxiliary tasks, such as recognition of phonemes, sentiment analysis, and recognition of named entities, LIEA allowed the model to learn specific structural and phonological features of Thai. This yielded measurable reductions in phoneme and classification loss, contributing to improved detection performance.

The model's sensitivity to language-specific tonal variations and syntax, learned through embedding alignment, indicates that contextual adaptation of pre-trained multilingual models is essential for hate speech detection in morphologically rich languages like Thai.

5.3. Proto-MAML and Low-Resource Learning

Although this study did not explore cross-lingual adaptation, the integration of Proto-MAML provided benefits for training stability and class representation during weak supervision. Using prototype-based class centers, the model maintained strong precision on the hate speech class even in a data-sparce environment. The few-shot adaptability of Proto-MAML facilitated efficient learning with limited noise-prone data.

5.4. Limitations and Cultural Sensitivity

While this approach achieved high accuracy, the study also revealed limitations inherent in low-resource NLP. Lexicon-based supervision may not fully capture the evolving, nuanced, or culturally coded nature of online hate speech. The static nature of manually curated lexicons limits the model's ability to adapt to new forms of expression or emerging slang.

Furthermore, even with embedding enrichment, certain benign terms were misclassified due to contextual ambiguity, such as the misinterpretation of the word "want" as a hate signal. This suggests the need for models that integrate deeper cultural context and pragmatic understanding beyond lexical or syntactic signals.

6. Conclusion

This study presents a weakly supervised, linguistically informed approach to hate speech detection in Thai, a low-resource language with limited annotated datasets. By combining sentiment-labeled data, a curated Thai toxicity lexicon, and auxiliary tasks in a linguistically enriched embedding space, we achieved high accuracy and macro F1 performance using minimal manual annotation. The integration of Proto-MAML further strengthened training efficiency and stability in a noisy, weak-label environment.

Our results demonstrate that weak supervision, when supported by linguistic alignment and regularization techniques, offers a practical and scalable pathway to build ro-

bust hate speech detection systems in under-resourced settings. The methodology presented here provides a replicable framework that can be adapted to other low-resource languages with similar structural challenges.

7. Future Work

Future research should focus on enhancing the cultural and linguistic adaptability of weakly supervised systems by:

- **Dynamic Lexicon Expansion:** Leveraging data mining and community contributions to update toxic lexicons with emerging hate speech terms.
- **Cultural Embedding Enrichment:** Training on localized corpora and integrating culturally sensitive attention mechanisms.
- **Advanced Meta-Learning Strategies:** Exploring task-aware and robust meta-learning approaches tailored to noisy supervision and context-sensitive tasks.
- **Longitudinal Analysis:** Monitoring evolving online discourse to inform adaptive model updates and detect changing trends in hate speech.

These extensions will support the development of more culturally aware, resilient, and inclusive NLP systems to detect harmful language in diverse online communities.

References

[1] Röttger P, Nozza D, Bianchi F, Hovy D. Data-Efficient Strategies for Expanding Hate Speech Detection into Under-Resourced Languages. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022 Dec; Abu Dhabi, United Arab Emirates. Stroudsburg (PA): Association for Computational Linguistics; 2022. p. 5674–91. Available from: https://aclanthology.org/2022.emnlp-main.383/

[2] Pookpanich P, Siriborvornratanakul T. Offensive language and hate speech detection using deep learning in football news live streaming chat on YouTube in Thailand. Social Network Analysis and Mining. 2024;14(18):1-14. doi:10.1007/s13278-023-01183-9

[3] Wikipedia contributors. List of languages by total number of speakers. Wikipedia, The Free Encyclopedia. Available from: `https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers`

[4] Kemp S. Digital 2024: Thailand. DataReportal – Global Digital Insights. Published January 2024. Available from: https://datareportal.com/reports/digital-2024-thailand

[5] Sirihattasak S, Komachi M, Ishikawa H. Annotation and classification of toxicity for Thai Twitter. In: Proceedings of the 2019 Conference on Computational Linguistics. 2019:243-256.

[6] Chormai P, Prasertsom P, Cheevaprawatdomrong J, Rutherford A. Syllable-based neural Thai word segmentation. In: Proceedings of the 28th International Conference on Computational Linguistics (COLING); 2020 Dec; Barcelona, Spain (online). Stroudsburg (PA): ACL; 2020. p. 4619–37. Available from: https://aclanthology.org/2020.coling-main.407/

[7] Ratner A, Bach S, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: Rapid Training Data Creation with Weak Supervision. arXiv preprint arXiv:1711.10160. Published November 27, 2017. Available from: https://arxiv.org/abs/1711.10160

[8] Suriyawongkul A, Chuangsuwanich E, Chormai P, Polpanumas C. PyThaiNLP: A Thai natural language processing library. Zenodo. 2019. doi:10.5281/zenodo.3457447

[9] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016:2818-2826. doi:10.1109/CVPR.2016.308

[10] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning. 2017:1126-1135.

[11] Snell J, Swersky K, Zemel RS. Prototypical Networks for Few-shot Learning. In: Advances in Neural Information Processing Systems (NeurIPS). 2017:4077-4087. https://arxiv.org/abs/1703.05175

[12] Maity K, Poornash AS, Bhattacharya S, et al. HateThaiSent: Sentiment-aided hate speech detection in Thai language. IEEE Transactions on Computational Social Systems. 2024. doi:10.1109/TCSS.2024.3376958

[13] Mnassri K, Zaghouani W. A survey on hate speech detection: From machine learning to deep learning. Journal of Natural Language Processing. 2024. doi:10.48550/arXiv.2401.12345

[14] Fortuna P, Nunes S. A survey on automatic detection of hate speech in text. ACM Computing Surveys. 2018;51(4):85. doi:10.1145/3232676

[15] Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. In: Proceedings of the 11th International Conference on Web and Social Media (ICWSM). 2017:512-515.

[16] Wang W, Sun C, Li Y, Wu J. Leveraging weak supervision for hate speech detection in multilingual settings. arXiv preprint arXiv:2301.01756. Published January 5, 2023. https://arxiv.org/abs/2301.01756

[17] Shu K, Mahudeswaran D, Wang S, Liu H. Cross-domain and weakly supervised hate speech detection: A survey. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2021:5485–5488. doi:10.1145/3340531.3412884

[18] Ruder S, Peters ME, Swayamdipta S, Wolf T. Transfer learning in natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorials. 2020:15-18. https://aclanthology.org/2020.emnlp-tutorials.4/

[19] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019:4171-4186. https://aclanthology.org/N19-1423/

[20] Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:8440-8451. doi:10.18653/v1/2020.acl-main.747

[21] Lample G, Conneau A. Cross-lingual language model pretraining. In: Advances in Neural Information Processing Systems. 2019:7059-7069.

[22] Chittaranjan G, Rajkumar R, Sitaram S. Word segmentation for languages without spaces. In: Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects. 2019:35-45.

[23] Piamsa N, Yang Y, Komachi M. A character-based Thai dependency parser. In: Proceedings of the 2018 International Conference on Computational Linguistics and Intelligent Text Processing. 2018:91-106.

[24] Awal MA, Karim MR, Islam MS, Rahman MS. HateMAML: Few-shot hate speech detection via meta-learning. In: Findings of the Association for Computational Linguistics: EMNLP. 2021:1234-1245.

[25] Mozafari M, Farahbakhsh R, Crespi N. Hate speech detection and racial bias mitigation in social media based on contextual embeddings. In: Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2020:409-416.

[26] Gupta S, Bhattacharyya P. Learning interpretable word embeddings via bidirectional alignment of embeddings with concepts. Knowledge-Based Systems. 2022;232:107471. doi:10.1016/j.knosys.2021.107471

[27] Park J, Lee J, Kim J, et al. LANGALIGN: Enhancing Non-English Language Models via Cross-Lingual Alignment at the Task Interface. arXiv preprint arXiv:2503.18603. Published March 2025. Available from: https://arxiv.org/abs/2503.18603

[28] Charoenporn T, et al. Survey on Thai NLP Language Resources and Tools. In: Proceedings of the 13th Language Resources and Evaluation Conference. European Language Resources Association; 2022:6461-6468.

[29] Maity K, Poornash AS, Bhattacharya S, et al. HateThaiSent: Sentiment-aided hate speech detection in Thai language. IEEE Trans Comput Soc Syst. 2024. doi:10.1109/TCSS.2024.3376958

[30] Pookpanich P, Siriborvornratanakul T. Offensive language and hate speech detection using deep learning in football news live streaming chat on YouTube in Thailand. Soc Netw Anal Min. 2024;14(18):1–14. doi:10.1007/s13278-023-01183-9

[31] Suriyawongkul A, Chuangsuwanich E, Chormai P, Polpanumas C. PyThaiNLP/wisesight-sentiment: First release [software]. Zenodo; 2019 Sep. doi:10.5281/zenodo.3457447. Available from: `https://doi.org/10.5281/zenodo.3457447`

[32] Sirihattasak S, Komachi M, Ishikawa H. Annotation and Classification of Toxicity for Thai Twitter 2019.