LGBT Cyberbullying Detection in Thai Language Utilizing Transformers-Based Algorithms

Vajratiya Vajrobol[1], Nitisha Aggarwal[1], Unmesh Shukla[1], Sanjeev Singh[1], Amit Pundir[2], Geetika Jain Saxena[2]

Institute of Informatics and Communication, University of Delhi, Delhi, India[1]
Maharaja Agrasen College, University of Delhi, Delhi, India[2]

Corresponding author: Amit Pundir, Email: amitpundir@mac.du.ac.in

Cyberbullying is a growing issue worldwide, particularly among minority groups such as the LGBT community. While research has shown that LGBT individuals are at a higher risk for experiencing cyberbullying, there is a lack of studies focusing on detecting cyberbullying directed towards this group in the context of the Thai language. This research aims to identify incidents of cyberbullying targeting the LGBT community in the Thai language on the social media platform, Twitter by using specific Thai derogatory keywords for data collection. In terms of model development for the detection of LGBT cyberbullying, deep learning-based algorithms, including CNN, LSTM, and transformer models have been applied to the Thai corpus to classify the cyberbullying messages. The results show that CNN performs better than other algorithms with 99.98% accuracy and 99.99% F1 score. The study aims to develop effective tools and strategies for detecting cyberbullying and ultimately contribute towards creating a better society that values diversity and inclusion of LGBT.

Keywords: Cyberbullying, LGBT, Thai language, Low-resource language

2023. In Satyasai Jagannath Nanda & Rajendra Prasad Yadav (eds.), *Data Science and Intelligent Computing Techniques*, 393–399. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.56155/978-81-955020-2-8-35

1. Introduction

As the world has transitioned into the digital age of information, social networking platforms on the internet have gained significant influence in society, particularly in Thailand with around 85% of internet users in 2021. As of the fourth quarter of 2022, 98.9 percent of the population who were 15 to 24 years old in Thailand used the Internet [1]. Research has indicated that Cyberbullying prevalence rates among adolescents vary widely worldwide, ranging from 10% to more than 70% in many studies. In Thailand, the prevalence of cyberbullying ranges from 39.0% to 70.7% [2]. In addition to Total Access Communication (DTAC) and media analytics firm Wisesight, two companies based in Thailand conducted a study on social media activity on platforms like Facebook, Instagram, and Twitter from 2018 to 2019. The study revealed that around 700,000 messages, or an average of 39 messages per minute, were associated with bullying or discrimination. These messages were then shared or retweeted, leading to an estimated total of 20 million instances. Of the messages, 36.4% were related to people's physical appearance, 31.8% to gender, 10.2% to opinions, and the remainder related to nationality, religion, traits, preferences, financial status, and family background [3].

As mentioned, cyberbullying can have different forms and targets, including those related to gender. Gender identity-based bullying refers to the use of derogatory language or behavior to harass or discriminate against someone based on their gender identity or sexual orientation. This can include spreading false information, using slurs, or making threatening comments online. Furthermore, studies have shown that cyberbullying can have negative impacts on the physical and mental health of young people. However, most of these studies have mainly focused on individuals who identify as heterosexual. Limited research has been conducted on sexual minority and gender non-conforming youth (i.e. LGBTQ), but the available studies indicate that this group is at a greater risk of experiencing cyberbullying compared to their heterosexual [4].

While Machine Learning (ML) can aid in cyberbullying detection, traditional ML approaches have limitations. They rely on predefined features and rules, which may not capture the complex and evolving nature of cyberbullying. Additionally, they struggle with generalization to new data. To address these challenges, advanced and adaptable ML techniques are needed to enhance cyberbullying detection. Therefore, the objective of this study is to identify incidents of cyberbullying targeting the LGBT community on Thai social media platforms. We collected data from Twitter using specific Thai derogatory keywords including labels associated with bullying the LGBT community and labelling as LGBT bullying label. To classify the cyberbullying messages, we employed a deep learning-based algorithm such as CNN (Convolution Neural Networks), LSTM (Long Short-Term Memory), Bi-GRU (Bidirectional Gated Recurrent Unit), and Transformer model-based approaches namely WangchanBERTa and TwHINBERT, known to be effective in various Thai text classification tasks such as sentiment analysis, emotion recognition, Natural language Inferences (NLI) [5].

The motivation of this research is to shed light on the issue of cyberbullying targeting the LGBT community on Thai social media platforms and to develop effective tools and strategies for detecting and addressing such incidents. By identifying and classifying such messages, this research aims to raise awareness about the prevalence and impact of cyberbullying on marginalized communities. Ultimately, the goal is to contribute towards creating a better society that values diversity and inclusion, where individuals can feel safe and respected regardless of their sexual orientation or gender identity. By reducing the spread of hate speech and promoting positive online interactions, this research could have a positive impact on the well-being and mental health of individuals who are often subjected to cyberbullying.

The subsequent sections of the research are structured as follows: Section 2 provides an overview of prior research conducted in the field of cyberbullying detection. In Section 3, we provide a comprehensive description of the dataset used in our study, along with the approach to detect cyberbullying. We then present a methodology for classifying LGBT cyberbullying. Moving on to Section 4, we present the experimental results obtained and provide a thorough analysis of the findings. Lastly, in Section 5, we conclude our work and outline potential avenues for future research.

2. Related Works

The prevalence of cyberbullying is increasing globally, with minority communities such as the LGBT group being at a higher risk of experiencing it. However, there is a dearth of studies focusing on the detection of cyberbullying targeting the LGBT community in the context of Thai language. This literature review seeks to examine previous research on cyberbullying detection. A recent study from [4] described that cyberbullying can result in negative effects on the physical and mental health of young people. However, most research studies have primarily focused on individuals who identify as heterosexual and cisgender. LGBTQ youth suggests that this group is at a greater risk of experiencing cyberbullying when compared to their heterosexual peers.

Focusing on machine learning and deep learning technology to detect social bullying. Kumar & Sachdeva (2022) proposed hybrid deep learning with Bi-GRU and self-attention to identify cyberbullying messages. The results have shown that it performs better than the conventional model with 09% increase in F1-score [6]. Neelakandan et al., (2022) developed the hybrid deep learning model based on the Salp swarm algorithm (SSA) and deep belief network (DBN), which is known as SSA-DBN model to detect cyberbullying and the results indicated that it outperforms other algorithms with 99.98 % accuracy [7].

Researcher also used hybrid models based on RNN or GRU to detect cyberbullying in Thai language on a social media dataset. In a 2020 study on cyberbullying detection, researchers used fine-tuning transformers and evaluated the xlm-Roberta-base model in three languages: English, Hindi, and Bengali. Initially, the model performed poorly in the Hindi sub-tasks. However, when trained jointly with different languages, its performance significantly improved, indicating the effectiveness of multi-language training in enhancing its[8]. Another study, applied transformer-based architectures for automated cyberbullying detection on Twitter comments on balanced and imbalanced datasets. The proposed framework outperforms baseline models and achieves an average F1-score of 95.92% on the balanced dataset and 87.51% on the imbalanced dataset. We also analyze cases where misclassification occurs, providing insights for improvement. Our models and code are publicly available [9].

In this study, the authors present a novel approach for cyberbullying detection using a transformer network-based word embedding technique. They utilise RoBERTa to generate word embeddings and employ Light Gradient Boosting Machine as a classifier. The proposed approach surpasses traditional machine learning algorithms like logistic regression and support vector machine, as well as deep learning models such as word-level convolutional neural networks (word CNN) and character convolutional neural networks with shortcuts (char CNNS), in terms of precision, recall, and F1-score. This highlights the effectiveness of the proposed approach in improving cyberbullying detection performance [10]. Recent study utilizes cyberBERT, BERT-based language model, for cyberbullying identification across different social media platforms. cyberBERTn model achieves state-of-the-art results on three real-world corpora: Formspring, Twitter, and Wikipedia, consisting of thousands of posts. The model outperforms existing attention-based deep neural network models, demonstrating significant improvements in cyberbullying detection [11]. In addition, Automated textual cyberbullying detection is a challenging task. Transfer learning through fine-tuning attention-based transformer language models, such as BERT and HateBERT, has achieved near state-of-the-art precision in identifying bullying-related text. This study examines whether these models learn cyberbullying attributes or syntactical features [12]. In terms of the detection of cyberbullying in Thai language, research collected and utilized several machine learning algorithms to classify textual tweets into four cyberbullying classes such as insult, sexual harassment, and threat, race and religion, and intelligence, appearance, and social status. The experiment identified that Logistic Regression (LR) performed best with an accuracy of 73.61% and F1 score of 73.89% [13].

3. Methods

3.1 Dataset

The dataset has been retrieved from Twitter using a Python package named "SNS scrape" (JustAnotherArchivist, n.d.). The tweets that are labeled as cyberbullying have been collected from Thai derogatory words related to LGBT such as "อีตุ๊ด", "อีเกย์", "ระเบิดถังขี้", "ขุดทอง", "อีผิดเพศ" as described in Table 1 and contains 10,072 records. While the non-cyberbullying dataset has been retrieved from the keyword "แฮปปี้" which means happy in Thai, the non-cyberbullying tweets contain 10,014 rows. The total raw dataset contains 20,086 records. However, the data has been pre-processed and cleaned by removing Thai punctuations and Thai stop words by using PythaiNLP package, and the whole dataset presents 20,047 records labeled as 9,984 rows and 10,063 rows with non-bullying and bullying classes, respectively. In terms of label encoding, the dataset has been encoded as 1 (bullying) and 0 (non-bullying), as illustrated in Table 2. The average text length of the whole dataset is about 13 words. In terms of dataset splitting, the dataset has been split to 80: 20 with training set and testing set, records with 16,037 and 4,010 respectively. The dataset can be downloaded through the GitHub repository [14].

Table 1. The description of keywords

Thai derogatory words related to LGBT	English Translation
อีตุ๊ด	Transgender in an abusive way
อีเกย์	Gay in an abusive way
ระเบิดถังขี้	Bombing poo
ขุดทอง	Digging Gold
อีผิดเพศ	Wrong gender in an abusive way

Table 2. An example of the dataset

Texts	English Translation
อิตุ๊ดขี้เสือก (F**cking Nosy parker transgender)	Bullying (1)
เช็คอินก่อนนอนทุกวันแฮปปี้ (Checking in before sleeping makes me happy.)	Non-Bullying (0)

3.2 Data-preprocessing

The data is preprocessed utilizing the PyThaiNLP library to tokenize and process Thai text data then removes certain punctuation marks and special characters from the text, and eliminates any Thai stop words from the text [15].

3.3 Text Classification Methods

3.3.1 CNN

CNNs are commonly used for image and text classification tasks. In the context of Thai text classification, CNNs are applied to the input Thai text to extract relevant features that can help in identifying the category or label of the text.

3.3.2 LSTM

LSTM models are designed to handle sequential data and can capture long-term dependencies in the data, making them effective for tasks such as Thai text classification. LSTMs are popular for NLP tasks due to their ability to handle variable-length sequences and avoid the vanishing gradient problem that can occur in traditional RNNs.

3.3.3 Bi-GRU

CNNs are commonly used for image and text classification tasks. In the context of Thai text classification, CNNs are applied to the input Thai text to extract relevant features that can help in identifying the category or label of the text

3.3.4 WangchangBERTa

WangchangBERTa is a language model that is based on the RoBERTa architecture, which was trained on a large, diverse dataset of social media posts, news articles, and other publicly available datasets. The training set was carefully selected and cleaned to eliminate duplicates, resulting in a total size of 78GB.WangchangBERTa outperforms several baseline models (NBSVM, CRF, and ULMFit) and multilingual models (XLMR and mBERT) on both sequence classification and token classification tasks in human-annotated, mono-lingual contexts.

3.3.5 TwHINBERT

The TwHINBERT, which is a network of diverse information on Twitter, has been trained on a massive dataset of over 7 billion tweets in more than 100 languages including Thai. The model has been tested on various tasks related to social recommendations and semantic understanding across multiple languages. TwHIN comprises around 200 million unique users, over 1 billion tweets, and more than 100 billion connections between them, all carefully selected and organized for analysis.

3.3.6 Evaluation Metrics

The standard of evaluation metrics has assessed the performance of models such as accuracy, recall, precision, and F1-score.

4. Results and Discussion

According to Table 3, the results presented show the performance of different algorithms in a classification task. It is evident that all the algorithms have achieved excellent performance across all evaluation metrics. Among them, the CNN algorithm stands out as it has achieved the highest values for accuracy, precision, recall, and F1-score. This indicates that the CNN algorithm has strong predictive power and is able to make accurate predictions with a high level of precision and recall.

Although the WangchanBERTa algorithm has slightly lower scores compared to the other algorithms, it still maintains a high level of performance. Despite the minor differences in scores, all the algorithms exhibit consistent and reliable performance in terms of accuracy, precision, recall, and F1-score. Moreover, the model "WangchanBERTa" has exhibited exceptional performance across multiple evaluation metrics. With an accuracy of 0.9935, it correctly classified 99.35% of the samples in the dataset. Additionally, the precision score of 0.9935 signifies that when the model predicts a sample as positive, it is accurate 99.35% of the time. Moreover, the model achieved a recall of 0.9935, indicating its ability to correctly identify 99.35% of the positive samples in the dataset.

These results demonstrate the effectiveness of the algorithms in making accurate predictions with high precision, recall, and F1-score. This suggests that the algorithms are reliable choices for classification tasks, where accurate identification and classification of instances are crucial. The superior performance of the CNN algorithm highlights its potential as a strong predictive model, while the WangchanBERTa algorithm, despite its slightly lower scores, remains a competitive choice for classification tasks. These findings emphasize the algorithms' effectiveness and ability to provide reliable predictions, making them valuable tools in various classification scenarios.

Table 3. The results of each deep learning method

Algorithms	Accuracy	Precision	Recall	F1-score
CNN	0.9998	1.0000	0.9998	0.9999
LSTM	0.9983	0.9977	0.9972	0.9975
Bi-GRU	0.9992	0.9993	0.9993	0.9993
WangchanBERTa	0.9935	0.9935	0.9935	0.9935
TwHINBERT	0.9992	0.9992	0.9992	0.9992

5. Conclusions

In conclusion, this study sheds light on the issue of cyberbullying targeting the LGBT community on Thai social media platforms. The results highlight the prevalence of cyberbullying and the negative impact it has on the physical and mental health of marginalized communities. By employing deep learning-based algorithms to classify cyberbullying messages, this research aims to raise awareness about the issue and develop effective tools and strategies for detecting and addressing such incidents. The results indicated that all the algorithms performed exceptionally well in the classification task. They achieved high accuracy, precision, recall, and F1-score values, commonly used metrics to evaluate the performance of classification models.

Among the algorithms, the CNN algorithm stood out as the top performer. It achieved the highest scores in all the metrics, including accuracy, precision, recall, and F1-score. This indicates that the CNN algorithm was able to make highly accurate predictions with a minimal number of false positives and false negatives. Its strong predictive power suggests that it effectively captured important patterns and features in the data.

The results highlight the effectiveness of the algorithms in accurately classifying instances. They provide a reliable and robust solution for classification tasks, with the CNN algorithm exhibiting robust predictive capabilities. These findings suggest that the algorithms can be trusted for making accurate predictions in various classification scenarios.

These findings can contribute towards creating a better society that values diversity and inclusion, where individuals can feel safe and respected regardless of their sexual orientation or gender identity.

By reducing the spread of hate speech and promoting positive online interactions, this research could have a positive impact on the well-being and mental health of individuals who are often subjected to cyberbullying. Overall, this study highlights the need for continued research and intervention efforts to address cyberbullying targeting the LGBT community in Thailand and promote a more tolerant and accepting society.

References

[1] STATISTA: Share of internet users in Thailand as of the 4th quarter of 2022, by age group.

[2] Thumronglaohapun, S., Maneeton, B., Maneeton, N., Limpiti, S., Manojai, N., Chaijaruwanich, J., Kummaraka, U., Kardkasem, R., Muangmool, T., Kawilapat, S., Juntaping, K., Traisathit, P., Srikummoon, P.: Awareness, perception and perpetration of cyberbullying by high school students and undergraduates in Thailand. PLoS One. 17, e0267702 (2022). https://doi.org/10.1371/journal.pone.0267702.

[3] NATION THAILAND: Dtac launches campaign to tackle cyberbullies.

[4] Abreu, R.L., Kenny, M.C.: Cyberbullying and LGBTQ Youth: A Systematic Literature Review and Recommendations for Prevention and Intervention. J. Child Adolesc. Trauma. 11, 81–97 (2018). https://doi.org/10.1007/s40653-017-0175-7.

[5] Harnmetta, P., Samanchuen, T.: Sentiment Analysis of Thai Stock Reviews Using Transformer Models. In: 2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE). pp. 1–6 (2022).

[6] Kumar, A., Sachdeva, N.: A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media. World Wide Web. 25, 1537–1550 (2022). https://doi.org/10.1007/s11280-021-00920-4.

[7] S, N., M, S., Chandrasekaran, S., K, M., Singh Pundir, A.K., R, S., Lingaiah, T.B.: Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media. Comput. Intell. Neurosci. 2022, 1–13 (2022). https://doi.org/10.1155/2022/2163458.

[8] Mishra, S., Prasad, S., Mishra, S.: Multilingual Joint Fine-tuning of Transformer models for identifying Trolling, Aggression and Cyberbullying at TRAC 2020. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying. pp. 120–125 (2020).

[9] Ahmed, T., Ivan, S., Kabir, M., Mahmud, H., Hasan, K.: Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying. Soc. Netw. Anal. Min. 12, 99 (2022).

[10] Pericherla, S., Ilavarasan, E.: Transformer network-based word embeddings approach for autonomous cyberbullying detection. Int. J. Intell. Unmanned Syst. (2021).

[11] Paul, S., Saha, S.: CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification. Multimed. Syst. 28, 1897–1904 (2022).

[12] Verma, K., Milosevic, T., Davis, B.: Can attention-based transformers explain or interpret cyberbullying detection? In: Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022). pp. 16–29 (2022).

[13] Semangern, T., Chaisitsak, W., Senivongse, T.: Identification of Risk of Cyberbullying from Social Network Messages. Presented at the (2019).

[14] Vajrobol, V.: Thai cyberbullying LGBT dataset.

[15] Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L., Chormai, P.: PyThaiNLP: Thai Natural Language Processing in Python, https://doi.org/10.5281/zenodo.7890372, (2023). https://doi.org/10.5281/zenodo.7890372